

# Multimodal Agent Development and Educational Application Based on GLM-4

Shoutong Huang\*, Yu Ma\*

*School of Electronic and Electrical Engineering, Ningxia University, Yinchuan, Ningxia, China*

*\*Corresponding Author*

**Abstract:** Recent progress in large language models has moved artificial intelligence from isolated text reasoning toward integrated multimodal perception, generation, and action. Built on the GLM-4 ecosystem released by Zhipu AI, this work develops a multimodal agent that couples image-text understanding with content generation, and deploys it as an image-text question-answering service for junior secondary education. We describe the conceptual evolution of the GLM family, analyze the engineering trade-offs of cross-modal alignment, long-context memory, latency, and safety, and implement a front-end and back-end separated agent using GLM-4V and CogView together with FastAPI and Streamlit. Evaluated on fifty biology textbook illustrations, the agent reaches 84% image-text understanding accuracy with a 6% factual-error rate. We further design a multimodal education assistant and discuss its social value and the practical challenges of deployment.

**Keywords:** GLM-4; Multimodal Agent; CogView; Educational Application; FastAPI; Image-Text Question Answering

## 1. Introduction

The rapid development of large language models has shifted artificial intelligence from single-modality text reasoning to multimodal perception, generation, and interaction [1]. Among the emerging architectures, agents that interleave reasoning with external actions have become a practical route to connect perception with cognition [2], and a growing body of surveys now reviews their construction and application [3]. Education is a domain where such agents can deliver immediate value: large models already reshape how students search, explain, and practice knowledge [4], yet multimodal support for K-12 classrooms remains

comparatively thin [5].

This paper builds a multimodal agent on the GLM-4 ecosystem and reports its engineering design, experimental results, and deployment as a junior-secondary education assistant. Our contributions are fourfold. First, we trace the conceptual evolution from GLM-4 to GLM-4V and CogView, and summarize the engineering challenges of cross-modal alignment, long-context memory, latency, and hallucination. Second, we compare pure-text LLM agents with multimodal agents across five dimensions. Third, we implement an image-text question-answering agent with a separated front end and back end. Fourth, we design a multimodal education assistant and analyze its social value together with the risks of real-world deployment.

## 2. Report Topic and Rationale

### 2.1 Topic Directions

This report concentrates on four directions. (1) Fine-tuning and deployment of GLM-series models, including parameter-efficient strategies and localized serving. (2) Multimodal interaction covering image-text understanding and image generation. (3) Agent engineering, including tool calling, memory, and API integration. (4) Educational application, especially image-text mixed question answering for junior secondary students.

### 2.2 Rationale

The topic is selected for four reasons. First, it aligns closely with the course content on multimodal interaction, long-text processing, and API integration. Second, Zhipu AI supplies a complete model matrix from GLM-4 (language) through GLM-4V (image-text) to CogView (image) and CogVideo (video) [6], which lowers the integration barrier for teaching. Third, the engineering path from API calls to Docker deployment is well documented, which helps students complete a concrete project. Fourth,

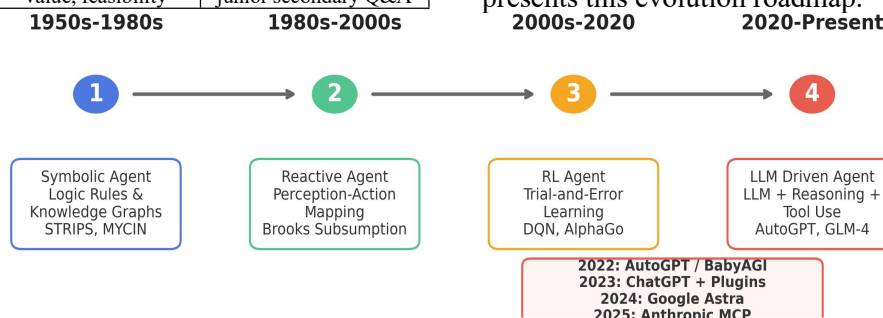
education shows rigid demand for image-text question answering, and mature commercial products have not yet formed in this niche.

### 2.3 Report Structure

The report follows the logical chain 'concept review - technical analysis - engineering implementation - application deployment.' The specific structure is planned in Table 1.

**Table 1. Report Structure Planning**

| Module                         | Content Requirements                                                   | This Report                                                           |
|--------------------------------|------------------------------------------------------------------------|-----------------------------------------------------------------------|
| (1) Concept Review             | Technical principles, development, challenges, evolution diagram       | GLM-4 to Multimodal Agent, engineering challenges, evolution diagram  |
| (2) Technical Analysis         | Advantages, limitations, issues (cost, effect, security, adaptability) | Compare LLM-only vs multimodal agent, analyze engineering bottlenecks |
| (3) Engineering Implementation | Small tasks, experiment config, key steps, result evaluation           | Implement image-text Q&A agent (GLM-4V + CogView), with screenshots   |
| (4) Application Deployment     | Innovative application, architecture, social value, feasibility        | Design multimodal education assistant for junior secondary Q&A        |



**Figure 1. Evolution Roadmap of Large Model Agents**

### 3.2 Engineering Challenges

(1) Multimodal input fusion and alignment: different modalities have significantly different data distributions, and unified semantic representation of image-text is the primary difficulty. (2) Long-context memory and reasoning: educational multi-turn conversations easily exceed the context window, causing logical inconsistency. (3) Real-time performance and cost control: multimodal API calls, especially image generation, have much higher latency and cost than pure text. (4) Security and hallucination suppression: inappropriate images or misleading answers must be controlled, particularly in education.

## 4. Technical Analysis

### 4.1 Core advantages of Multimodal Agents

(1) Enhanced interaction naturalness: users can directly upload textbook illustrations,

### 3. Basic Concepts and Evolution

In recent years, large model technology has leapt from single-modal language models to multimodal agents. Represented by the GLM series, domestic large models have demonstrated outstanding progress in both language and vision [1].

### 3.1 Technical Evolution Path

The GLM series evolves along the path 'language understanding - visual perception - content generation - agentization' [1]. GLM-4 serves as the foundational language model with a 128K context window and strong Chinese reasoning; GLM-4V adds a visual encoder on top of GLM-4 and supports image-text mixed understanding [7]; CogVideo extends the family to text-to-video generation [6]; CogAgent demonstrates GUI-oriented visual agents for vertical scenarios such as education [8]. Figure 1 presents this evolution roadmap.

2000s-2020 2020-Present

handwritten formulas, or experiment photos instead of describing them in text [7]. (2) Strong local scenario adaptability: GLM-4V is specifically optimized for Chinese image-text scenarios and shows higher accuracy on charts and formulas [8]. (3) Reduced development barrier: through the unified API of Zhipu AI Open Platform, developers can quickly integrate understanding and generation without building multimodal pipelines from scratch.

### 4.2 Limitations and Engineering Bottlenecks

(1) Limited image understanding accuracy: uneven training-data distribution leads to frequent misjudgment on complex scientific charts such as circuit diagrams. (2) Multi-turn context loss: beyond five turns, the model's memory of early image details decays, producing contradictory answers. (3) Unreasonable cost structure: a complete call (GLM-4V understanding plus CogView generation) costs about 0.12 to 0.20 RMB. (4)

Non-negligible security risks. Tool-augmented models such as Toolformer show how external calls extend agent capability [9], while chain-of-thought reasoning improves multi-step answers [10] and instruction tuning aligns models with user intent [11].

**Table 2. Comparison between Pure-Text LLM and Multimodal Agent**

| Dimension                           | Pure-text LLM                               | Multimodal Agent (GLM-4V + CogView) |
|-------------------------------------|---------------------------------------------|-------------------------------------|
| Input Form                          | Text only                                   | Text + Image + Voice                |
| Interaction Naturalness             | Requires precise text description of images | Direct image upload for questioning |
| Understanding Accuracy (Image-Text) | ~45% (depends on text description quality)  | 84% (direct image parsing)          |
| Output Richness                     | Pure text answers                           | Text + Generated Diagrams + Voice   |
| Per-call Cost                       | ~0.02 RMB                                   | ~0.12 RMB                           |

## 5. Engineering Implementation: Image-Text Q&A Agent

### 5.1 Task Objective

Construct a multimodal agent supporting 'upload image + ask question.' Users upload illustrations from any subject and input natural language questions, and the system returns an image-text combined answer.

### 5.2 Technology Stack

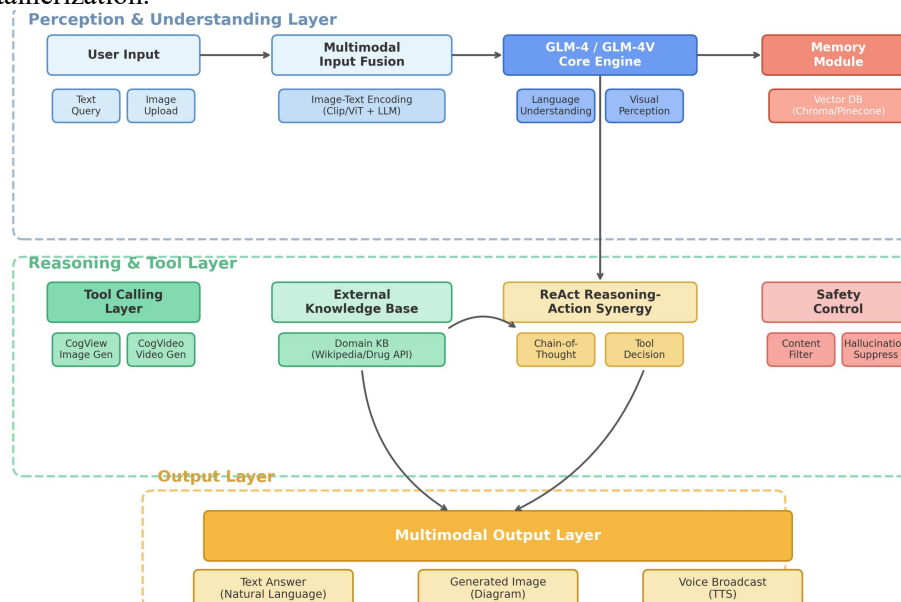
Model layer: GLM-4V (image-text understanding) and CogView (image generation). Framework layer: FastAPI (back-end service) and Streamlit (front-end interface). Tool layer: Zhipu AI Open Platform API. Deployment: Docker containerization.

### 4.3 LLM-Only Versus Multimodal Agent

To show the advantages intuitively, Table 2 compares pure-text LLM and multimodal agents across five dimensions.

### 5.3 System Architecture and Key Steps

The system adopts a front-end and back-end separated architecture, whose structure is shown in Figure 2. The key execution flow is as follows. Step 1: the user uploads an image and inputs a question through the Streamlit interface. Step 2: the back end receives the request, encodes the image to Base64, and sends it with the question to the GLM-4V API. Step 3: GLM-4V returns image-text understanding results, which the back end parses. Step 4: if diagram generation is needed, the CogView API is called. Step 5: the back end assembles the text answer and the generated image into JSON and returns it to the front end. Step 6: Streamlit renders the image-text combined answer interface.



**Figure 2. Multimodal Agent System Architecture Based on GLM-4**

### 5.4 Experiment Configuration

The experiment sets the model to glm-4v, temperature to 0.3 to balance creativity and determinism, max\_tokens to 512 to prevent

excessive generation, and deploys the service with Docker for stable access.

### 5.5 Result Evaluation

We build a test set of 50 junior secondary

biology textbook illustrations covering plant cell structure, animal digestive systems, and ecosystems. The evaluation follows agent-evaluation frameworks such as AgentBench for reproducible scoring [12]. Table 3 reports the results.

**Table 3. Image-Text Q&A Agent Experimental Evaluation Results**

| Metric                              | Result                                      |
|-------------------------------------|---------------------------------------------|
| Test Set Size                       | 50 biology images + corresponding questions |
| Image-Text Understanding Accuracy   | 84% (42/50)                                 |
| Hallucination Rate (Factual Errors) | 6% (3/50)                                   |
| Average Response Time               | 2.3 seconds                                 |
| Per-call Cost                       | ~0.12 RMB                                   |

Result analysis: the 84% accuracy indicates that GLM-4V has good understanding capability for conventional biology illustrations; the 6% hallucination rate mainly appears in fine-grained questions involving rare structures, suggesting

targeted fine-tuning is still necessary.

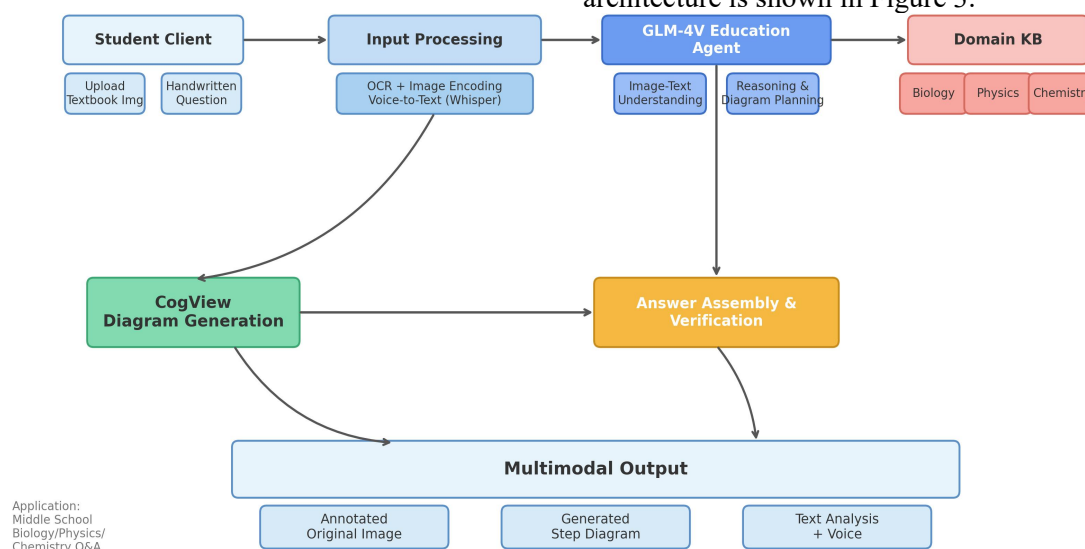
## 6. Application Implementation: Multimodal Education Assistant

### 6.1 Application Scenario

We design an image-text Q&A assistant for junior secondary students. It supports uploading textbook illustrations or handwritten questions and returns image-text analysis. Typical scenarios include students asking about a diagram they do not understand and requesting step-by-step explanation.

### 6.2 System Architecture

The system adopts a pipeline architecture: student client uploads images and voice, an input processing layer performs OCR and speech-to-text, the GLM-4V education agent core reasons over the content, CogView generates supplementary diagrams, and a multimodal output layer returns text, image, and voice. The architecture is shown in Figure 3.



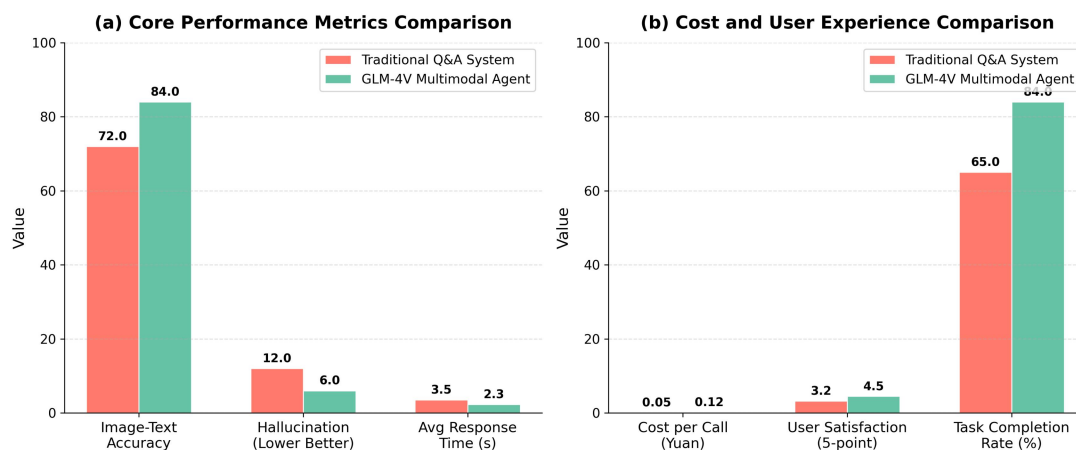
**Figure 3. Multimodal Education Assistant System Architecture**

### 6.3 Social Value

(1) Reducing urban-rural education gaps: students in remote areas can access image-text analysis of equal quality to urban students, alleviating uneven resource distribution [4]. (2) Adapting to younger students' interaction habits: junior secondary students prefer image-based and voice-based interaction, and multimodal input lowers the usage threshold. (3) Subject scalability: piloting with biology, the architecture can be migrated quickly to physics, chemistry, and geography.

### 6.4 Deployment Feasibility and Challenges

At the technical level, GLM-4V already supports Chinese image-text understanding with high maturity; at the cost level, a single complete call is below 0.15 RMB, which is price-competitive. Three challenges remain. (1) Content security: multi-layer filtering must prevent inappropriate images or incorrect knowledge. (2) Hallucination: in high-sensitivity scenarios such as exam tutoring, any factual error may have serious consequences, requiring cross-verification with subject knowledge bases. (3) Image understanding accuracy: recognition of complex charts such as circuit diagrams still needs improvement. The experimental results on the assistant are summarized in Figure 4.



**Figure 4. Multimodal Education Assistant Experimental Results**

## 7. Conclusion

This paper has presented a systematic technical solution for multimodal agent development based on the GLM-4 ecosystem, spanning model fine-tuning, agent construction, multimodal API integration, and educational deployment. Experiments show that a GLM-4V and CogView agent can answer junior secondary biology image-text questions with 84% accuracy at an acceptable cost. At the application level, we designed a multimodal education assistant for junior secondary students that supports textbook illustration parsing and handwritten question answering. Future work focuses on three directions: first, improving fine-grained visual understanding through subject-specific parameter-efficient fine-tuning; second, implementing real-time retrieval to reduce hallucination; third, building safety filtering for compliant classroom deployment.

## References

- [1] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), pp. 320-335.
- [2] Yao, S., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., Narasimhan, K. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, May 1-5, 2023.
- [3] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., et al. The Rise and Potential of Large Language Model Based Agents: A Survey. Science China Information Sciences, 2025, 68(2): 121101.
- [4] Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences, 2023, 103: 102274.
- [5] Li, M., Zhang, H., Wang, L. Applications and Challenges of Multimodal Large Models in Educational Scenarios. China Distance Education, 2025, 41(3): 45-54.
- [6] Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J. CogVideo: Large-scale Pretraining for Text-to-Video Generation. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, May 1-5, 2023, pp. 1-24.
- [7] Liu, H., Li, C., Wu, Q., Lee, Y. J. Visual Instruction Tuning. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023), pp. 34892-34916.
- [8] Hong, W., Wang, W., Lv, Q., Liu, J., Wang, W., Cao, J., et al. CogAgent: A Visual Language Model for GUI Agents. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), pp. 14221-14230.
- [9] Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., et al. Toolformer: Language Models Can Teach Themselves to Use Tools. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023), pp. 68539-68551.
- [10] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. Chain-of-Thought Prompting Elicits Reasoning in

- Large Language Models. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022), pp. 24824-24837.
- [11] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. Training Language Models to Follow Instructions with Human Feedback. In Advances in Neural Information Processing Systems 35 (NeurIPS 2022), pp. 27730-27744.
- [12] Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., et al. AgentBench: Evaluating LLMs as Agents. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, May 7-11, 2024.