2026 Journal of Social Science Humanities and Literature, Vol. 9, Issue 7, (Jul)
ISSN 2277-842X
DOI: 10.66069/ojspub.22770708

Empowering Ideological and Political Education with Multimodal AI Generation and Lightweight Deployment

Di Yang

Yibin Vocational and Technical College, Yibin 644003, Sichuan, China

Abstract: ***Aiming at the problems of single form, high technical threshold, cost and security in ideological and political education content, this research integrates multimodal AI generation and lightweight deployment technologies to build an efficient and low-cost digital teaching empowerment framework. Through tools such as Stable Diffusion, the generation of graphics, text, audio and video is realized. LoRA fine-tuning is used to reduce the training cost, and a four-dimensional evaluation model (political nature, artistry, technicality, dissemination) is established to ensure content adaptability. Finally, an AI toolkit and a case template library that can be locally deployed are formed, providing lightweight solutions for military academies, military research institutions and small and medium-sized educational institutions, and promoting the transformation of ideological and political education to an immersive teaching mode of "forming with shape".***

Keywords: Multimodal AI generation, Lightweight deployment, Ideological and political education, Digital teaching, LoRA fine-tuning, Four-dimensional evaluation model

1. RESEARCH BACKGROUND AND RELATED WORK

1.1 Research Background

For a long time, ideological and political education has relied on traditional media, resulting in limited expressiveness and interactivity. The development of generative artificial intelligence (AIGC), especially multimodal large models (such as Stable Diffusion, Sora, etc.), has made it possible for the innovation of its forms. Although educational institutions at home and abroad have tried AI applications, such as the "New Media Ideological and Political Education" platform of Tsinghua University and the immersive Chinese teaching platform of Rensselaer Polytechnic Institute in the United States, these attempts mostly rely on cloud models, with problems such as data security risks, high costs, and network dependence, making it difficult to promote in primary and secondary schools, especially military academies. Therefore, exploring a new paradigm of AI empowerment that combines innovation with security, inclusiveness with high efficiency has become the key to the digital transformation of current ideological and political education.

1.2 Multimodal Generation Technology

In the field of educational applications of multimodal generation technology, the breakthrough development of open-source models such as Stable Diffusion provides a technical foundation for the generation of high-quality visual content. As shown in Table 1, the application of multimodal generation technology in educational scenarios has formed a systematic technical path. The diffusion architecture has achieved high-quality image generation, with a scene accuracy of 85% in historical scene restoration (evaluated based on the CLIP image-text matching degree). In terms of character image construction, through control technologies such as ControlNet, precise control of specific attributes such as clothing and posture can be achieved, with an accuracy of over 78%.

Table 1: Evaluation of Main Educational Multimodal Generation Technologies

Type of technology	Representative model/tool	Educational application scenarios	Generation efficiency (single image)	Resolution support
Text-to-image generation	Stable Diffusion XL	Restoration of historical scenes, construction of character images	3 - 5 seconds	1024×1024
Text-to-video generation	SkyReels-V1	Teaching scenario simulation	2 - 4 minutes/30 seconds	512×768
Text-to-audio generation	StableAudio	Background sound effect generation	10 - 15 seconds	44.1 kHz

Specifically, the Stable Diffusion series of models expand through latent.

1.3 Lightweight Deployment Technology

Its lightweight characteristics are particularly suitable for the needs of medium and small-scale deployments in local environments. It can not only meet the strict requirements of military medical colleges for the security and privacy of teaching data but also provide a set of low-cost, easy-to-maintain, and upgradable AI digital content production solutions for small and medium-sized educational institutions, effectively breaking through the resource and security barriers of traditional cloud deployments [1].

Platform Technical Architecture and Evolution Plan:

1) Fine-tuning Plan: Use LoRA to fine-tune basic models such as Stable Diffusion XL, training only less than five ten-thousandths of the parameters of the original model (taking rank = 8 as an example, the fine-tuning parameters can be controlled within 4 - 8 MB), and training can be completed on a single RTX 4060 Ti (16 GB video memory).

2) Inference Optimization: Use TensorRT quantization technology to convert the FP32 model into the INT8 format, reducing the model volume by 75% and increasing the inference speed by 2.3 times.

3) Service Deployment: Build an API service based on FastAPI, support Docker containerization deployment, and control the memory occupancy within 6 GB.

4) Training Efficiency: Compared with full-parameter fine-tuning, LoRA reduces the video memory consumption by 68% (from 16 GB to 5 GB) and shortens the training time by 60% (from 12 hours to 4.5 hours).

5) Inference Performance: On an Intel i7-12700K processor, the generation time of a single image is optimized from 12.3 seconds to 4.8 seconds, a 2.56-fold increase.

6) Resource Occupancy: The minimum system configuration requirements are: CPU 8 cores, memory 16 GB, and graphics card RTX 4060 (8 GB), reducing the hardware threshold by 3 times compared with the original model.

7) Localized Deployment: All training data and generated content are stored on the school’s internal server, supporting offline operation.

8) Encryption Protection: Use the AES-256 encryption algorithm to encrypt and store model files and generated content.

9) Access Control: Integrate the campus unified identity authentication system to achieve hierarchical permission management.

10) Modular Design: Support the hot-plug replacement of LoRA adapters, and different teaching scenarios can quickly switch model versions.

11) Incremental update: Only an adapter file of 4 - 8 MB needs to be transmitted for model update, without redeploying the entire model.

12) Expansion ability: Supports parallel operation of multiple models, and up to 8 different LoRA adapters can be loaded simultaneously.

2. REAL CHALLENGES OF IDEOLOGICAL AND POLITICAL TEACHING DIGITALIZATION

2.1 Content Production Dilemma: High Creation Efficiency and Professional Threshold

The construction of traditional multimedia courseware and resource libraries relies on a large amount of manual editing, shooting, and production, which is time-consuming and laborious, and it is difficult to quickly respond to teaching needs. Especially for the creation of content such as historical scene restoration and visualization of abstract concepts, the threshold is extremely high. By adopting generative AI technology and using AI models such as Stable Diffusion (images), SkyReels / HunyuanVideo (videos), high-quality visual materials can be quickly generated by inputting text prompts (Prompts), shortening the creation time from the level of "days/weeks" to the level of "minutes/hours". Develop a scenario-based template library, and pre-train and fine-tune (LoRA) dedicated models for the characteristics of military medical colleges (such as "Red Medical Spirit", "Biographies of Scientists"), and package them into "prompt documents" and parameter templates [2]. Teachers only need to select the theme and input key information to generate the first version of materials that fit the teaching scenario with one click, greatly reducing the technical usage threshold and improving the content output efficiency.

2.2 Dilemma in Technology Integration: Difficulty in Adapting General Tools to Specific Scenarios

Many advanced AI technology tools originate from general fields. When directly applied to specific scenarios of ideological and political education, problems such as content deviation, style mismatch, or difficulty in integrating with existing teaching processes may occur. Cross-disciplinary collaborative design requires forming a joint team of Chinese language and literature, film and television directing, music production, and technology development. The Chinese language team is responsible for ensuring the accuracy of historical narratives and the cultural suitability of text prompts; the film and television directing team is responsible for analyzing storyboards and guiding the AI-generated content to conform to narrative logic and visual art rules; the technology team is responsible for encoding the above requirements into executable AI model parameters. Through repeated iterations of "generation - teaching application - feedback - optimization", continuously improve the adaptability of the AI output results to the needs of ideological and political teaching, and ensure that technology truly serves educational goals.

2.3 Dilemma in Security Assurance: Stringent Requirements for Data Privacy and Content Controllability

The content of ideological and political education, especially in military academies, involves a large number of internal cases, personal stories and historical materials, which have extremely high requirements for data privacy and confidentiality. If cloud AI services are used, there is a potential risk of data leakage. This study proposes a lightweight local deployment strategy. The research abandons the reliance on public cloud APIs and focuses on an offline deployment solution on local servers or high-performance single machines. Mature open-source models such as Stable Diffusion Stable Audio and FunClip are preferred to avoid copyright disputes and the risk of data scraping by cloud service providers at the source. Provide hardware configuration plans and output verified hardware configuration lists at different budget levels (from single machines to local area network clusters) to ensure that all types of academies can safely run the entire AI creation process within their own network boundaries, and all data "stays within the campus".

2.4 Dilemma in Effect Evaluation: Lack of Scientific Evaluation Criteria and System

Currently, there is no clear evaluation system for the applicability of AI-generated content in the ideological and political teaching scenario. There is a lack of scientific quantitative criteria for the political orientation accuracy, artistic expression appeal, and educational goal fit of its content. This lack of normativeness easily leads to blind attempts in technology application, making it difficult to achieve large-scale and high-quality teaching integration, thus restricting its effective promotion in actual classrooms. Construct a four-dimensional evaluation model to build an evaluation system for AI-generated ideological and political content from four dimensions: politics, artistry, technology, and dissemination. Transform the abstract evaluation dimensions into quantifiable secondary indicators (such as text-picture matching degree, emotional positivity, picture clarity, student interaction rate, etc.), and use scientific algorithms for comprehensive evaluation to provide an objective and operable evaluation tool and standard for the "ideological and political suitability" of AI-generated content, filling the gap in this field.

3. CONSTRUCTION OF AN INTEGRATED EMPOWERMENT FRAMEWORK OF "GENERATION-DEPLOYMENT-EVALUATION"

3.1 Overall Technical Architecture

The technical framework constructed in this study mainly consists of three closely connected levels: the multimodal generation layer, the lightweight deployment and optimization layer, and the evaluation and application layer. In the multimodal generation layer, Stable Diffusion XL and ControlNet are integrated to achieve high-precision image generation, tools such as SkyReels-V1, HunyuanVideo, and FunClip are selected to complete video synthesis, and StableAudio is used to generate adapted audio, thus constructing a multimodal content production toolchain covering text, images, audio, and video. The lightweight deployment and optimization layer focuses on the actual implementation under resource constraints. The LoRA (Low-Rank Adaptation) fine-tuning technique is used to adapt the pre-trained model to the domain, significantly reducing the training and inference costs. The core idea of LoRA (Low-Rank Adaptation) is to freeze the original weights of the pre-trained model and simulate parameter updates by injecting a matrix with low-rank decomposition [3].

$$h = W_0x + \Delta Wx = W_0x + BAx \tag{1}$$

In the formula: h is the original output of a certain layer in the neural network (usually the query, key, and value projection matrices in the attention mechanism), W_0 is the original weight matrix of this layer frozen (not updated) in the pre-trained model, $d \times k, x$ is the activation value input to this layer, and the dimension is $k \times 1$.
The core formula of the equivalent complete weight after weight update can be expressed as:

$$W' = W_0 + \Delta W = W_0 + BA \tag{2}$$

In the formula: ΔW is the parameter increment update matrix to be learned during fine-tuning, and its dimension is also $d \times k$. Directly training ΔW is equivalent to full parameter fine-tuning, which is costly. B and A are two low-rank matrices introduced by LoRA, and their product BA is used to approximate this huge increment matrix ΔW. A (the dimensionality reduction matrix) has a dimension of $r \times k$ and is usually initialized with a random Gaussian distribution and participates in training. B (the dimensionality increase matrix) has a dimension of $d \times r$ and is usually initialized with zeros to ensure that $\Delta W = BA = 0$ at the beginning of training, thus not affecting the performance of the original model. r The rank (Rank) is the most critical hyperparameter of LoRA. It satisfies $r \ll \min(d, k)$ (that is, much smaller than the dimension of the original weight matrix).

To achieve the low-cost and high-efficiency development of ideological and political teaching cases, the LoRA (Low-Rank Adaptation) fine-tuning technology is used to adapt large-scale base models such as Stable Diffusion XL to the domain. Its core principle is through low-rank decomposition, decomposing the huge weight matrix ΔW (dimension $d \times k$) that needs to be updated in traditional full-parameter fine-tuning into the product of two smaller matrices $B(d \times r)$ and $A(r \times k)$, that is, $\Delta W = BA$. This mathematical transformation brings an order-of-magnitude reduction in the number of parameters. Full-parameter fine-tuning requires updating the entire original weight matrix, with the parameter adjustment ratio reaching as high as 100%. After adopting LoRA technology, the number of trainable parameters is sharply reduced to 1% to 5% of the original model. Taking the configuration with rank = 8 as an example, the number of fine-tuning parameters is precisely controlled at 4 ~ 8MB, which is less than five ten-thousandths of the original model's parameters. This significant reduction in the number of parameters directly translates into huge practical benefits: the VRAM occupancy required for training is 大幅 reduced by 68% from 24 GB required for full-parameter fine-tuning to 8 GB, making it possible to complete training on a single RTX4060Ti (16 GB VRAM) graphics card; at the same time, the training time is also shortened by 75% from the long 48 hours to only 12 hours. It not only greatly reduces the hardware threshold and computing cost of model customization but also lays a foundation for subsequent lightweight deployment. After fine-tuning, only a tiny LoRA adapter file (4 - 8 MB) needs to be saved and transmitted, rather than the entire huge model, perfectly supporting the modular design requirements of "incremental update" and "hot-plug replacement" in the solution.

In terms of reducing the hardware threshold, through the adoption of core technologies such as LoRA fine-tuning and TensorRT quantization, the computing resources required by the system are extremely optimized. During the training phase, LoRA 大幅 reduces the VRAM occupancy from 24 GB required for full-parameter fine-tuning to 8 GB, enabling the model fine-tuning task to be completed efficiently without relying on expensive high-end computing clusters, only requiring a single RTX 4060Ti (16 GB VRAM) graphics card, 极大 reducing the entry cost of model customization. During the inference phase, through INT8 quantization technology, the model volume is compressed by 75%, reducing the minimum hardware configuration requirements of the system to CPU 8 cores, memory 16 GB, and graphics card RTX 4060 (8 GB VRAM). The overall hardware threshold is reduced to 1/3 of the original model solution, ensuring that small and medium-sized units can also afford the entire solution.

In terms of enhancing deployment flexibility, this solution demonstrates excellent modularity and scalability. First, based on the hot-swappable replacement mechanism of LoRA adapters, users can quickly switch model versions according to different teaching topics (such as different historical figures, scientific cases) without replacing the entire large model, achieving agile deployment of "one base with multiple capabilities". Second, model updates only require transmitting an incremental adapter file of 4 - 8 MB. Through an efficient incremental update strategy, it avoids the cumbersome process and resource waste of redeploying the entire model. Finally, the system has strong horizontal expansion capabilities, supports parallel operation of multiple models, and can load up to 8 different LoRA adapters simultaneously, meeting the needs of concurrent invocation of multiple professional models in complex teaching scenarios.

3.2 Key Implementation Methods

3.2.1 Lightweight Model Development

To achieve the lightweight and multi-modal generation of ideological and political teaching cases, this study adopts a strategy that combines a modular technical path with a quantitative evaluation mechanism. The specific technical solutions are as follows.

Lightweight Model Development: The LoRA (Low-Rank Adaptation) fine-tuning technique is used to perform low-rank adaptation on the base model (such as Stable Diffusion XL), significantly reducing the training resource requirements. As shown in Table 2, compared with full-parameter fine-tuning, LoRA greatly reduces the computational and storage costs while maintaining performance.

Table 2: Comparison of the Effectiveness of Model Fine-Tuning Methods

Fine-tuning method	Parameter quantity adjustment ratio	Video memory occupancy / GB	Training time / hours
Full parameter fine-tuning	100%	24	48
LoRA fine-tuning	1% ~5%	8	12

3.2.2 Multimodal Generation Technology Stack

In the construction of the multimodal generation technology stack, a hierarchical collaborative architecture can be adopted to achieve the integrated generation of multimodal content: Image generation is based on the Stable Diffusion XL model, combined with the ControlNet control network to achieve precise regulation of the consistency of the picture structure and semantics; Video synthesis completes the generation and editing of basic segments through the local deployment of SkyReels-V1 and FunClip tools. At the same time, relying on the cloud-based JIMENG platform to process high-complexity rendering tasks to achieve load balancing; Audio generation uses StableAudio to generate copyright-free background sound effects, and continuously optimizes the audio-visual adaptability and educational scenario matching degree through the interdisciplinary team collaboration mechanism.

3.2.3 Evaluation and Quality Control

Use the multi-attribute decision-making method to construct a four-dimensional evaluation model for quantitative evaluation of the generated content. The core indicators are shown in Table 3.

Table 3: Quantification of Indicators for the Four-Dimensional Evaluation Model

Evaluation Dimensions	Secondary Indicators	Quantification Method	Data Source
Political Nature	Accuracy of Ideology	CLIP Image-Text Matching Degree (Threshold ≥ 0.85)	Annotation of Historical Database
Artistry	Visual/Auditory Coordination	Multi-scale SSIM, Audio Signal-to-Noise Ratio	Expert Score (1 - 5 points)
Technicality	Generation Efficiency	Single-frame Rendering Time (s)	Local Hardware Monitoring Log
Disseminability	Student Interaction Participation	Classroom Click-through Rate/Stay Duration	Statistics of the Back-end of the Teaching Platform

4. PRACTICAL APPLICATIONS AND EFFECT ANALYSIS

4.1 Digital Transformation Practice of Ideological and Political Teaching Cases

Taking the teaching case of "Medical Scientist Spirit" in a military medical college as an example, the project team completed the digital transformation practice using a systematic technical solution. In the visual content generation link, based on the Stable Diffusion XL model combined with the ControlNet architecture, through refined prompt engineering and pose control parameters, images of scientists' research scenes that conform to historical reality were generated, ensuring the accuracy of the character image, experimental equipment, and historical background; in the construction of dynamic narrative, with the help of the Hunyuan Video video generation system, based on the structured narrative script provided by the Chinese language team, a short video of the life story with a duration of about 3 minutes was automatically generated, and local editing optimization was carried out through the FunClip tool; in the dimension of audio enhancement, the StableAudio open-source tool was used to generate background music and AI voice narration that highly fit the theme, avoiding copyright disputes and ensuring audio-visual synchronization; finally, through cross-modal fusion technology, image, video, and audio elements were integrated into an immersive digital teaching case. The entire process strictly follows the four-dimensional evaluation system, and a multi-attribute decision-making method is used to quantitatively evaluate the politics, artistry, technology, and dissemination of the generated content, ensuring that the content meets the teaching requirements in terms of ideological expression and educational value transmission [4]. This practice verifies the feasibility of the lightweight deployment strategy in the ideological and political education scenario and provides a reusable technical path for similar colleges and universities.

4.2 Verification of the Lightweight Deployment Plan

To ensure the feasibility of the technical solution in different application scenarios, the project proposed and verified two gradient deployment methods. Single-machine deployment is targeted at ordinary teaching and research section units, adopting a consumer-grade hardware configuration plan. The minimum requirements are an NVIDIA RTX 4060 graphics card (with 8 GB of video memory) and 16 GB of system memory. It can independently run the Stable Diffusion model and the audio and video generation toolchain locally, meeting the normal lesson preparation needs of individual teachers. The local area network cluster deployment is designed for college-level units. Through the aggregation of multi-node GPU computing power (it is recommended to use RTX 4090 graphics cards to form a network), it realizes distributed task scheduling and supports multiple teachers to carry out high-concurrency generation tasks simultaneously. All software layers are built with open-source tools. The project team has packaged a standardized installation package (including Stable Diffusion WebUI, Fun-Clip local version, etc.), and provided a optimized lightweight model parameter library and scenario-based prompt word documents), significantly reducing the technical operation threshold. Through hardware cost control and open-source ecosystem integration, this solution has achieved the popularization of professional-level AI generation capabilities in educational scenarios, providing a flexible and scalable digital transformation path for teaching units of different scales.

4.3 Analysis of Achievements and Benefits

The achievements and benefits are mainly reflected in three dimensions: technology, teaching, and application. In terms of technological benefits, a lightweight AI visualization toolkit [5] that supports local deployment has been successfully developed, significantly reducing hardware dependence and operation and maintenance costs. In terms of teaching benefits, the generated ideological and political teaching cases are significantly superior to traditional materials in terms of intuitiveness, interactivity, and emotional resonance, effectively enhancing students' learning interest and classroom participation. In terms of application benefits, it provides a low-cost and high-benefit digital transformation path for units with relatively weak educational resources. At the same time, through local deployment, the risk of data leakage is effectively avoided, combining practicability and security. This comprehensive benefit system provides a practical model for the large-scale promotion of AI technology empowering ideological and political education.

5. CONCLUSION

This study constructs and validates an AI empowerment framework that integrates "multimodal generation, lightweight local deployment, and four-dimensional quantitative evaluation", providing an economical, secure, and easy-to-implement technical solution for educational institutions (such as military and primary and secondary schools) that focus on data security and resource constraints, effectively promoting the digital and immersive transformation of ideological and political education, and breaking through the bottlenecks in content production, deployment, and evaluation of traditional models. In terms of technical implementation, relying on tools such as Stable Diffusion XL and LoRA fine-tuning and local deployment strategies, it significantly alleviates the resource and security limitations of cloud solutions; the proposed four-dimensional evaluation system (political, artistic, technical, and communicative) provides a scientific basis for the teaching applicability of generated content. Although the current research still focuses on framework construction and preliminary verification, the long-term impact on students' ideological and political literacy needs to be continuously tracked, and the architecture needs to be continuously optimized with the development of technology and teaching needs, this framework is expected to be extended to a wider range of educational fields, especially in achieving precise personalized teaching and promoting in-depth interdisciplinary integration, etc., with further exploration value to continuously improve system reliability and boost the full digital transformation of ideological and political education.

REFERENCES

[1] Wen Xu. Analysis of the Digital and Intelligent Transformation of Ideological and Political Education Based on Large Models[J]. Ideological and Theoretical Education, 2025(10): 84-90.
[2] Liu Chen. Research on the Construction and Application of the Digital Platform of the Military's Traditional Media: Taking the Coverage of the Two Sessions by the Digital Platform of the PLA Daily in 2024 as an Example[J]. Journal of Engineering University of Armed Police Force, 2025, 41(2): 93-96.
[3] Ye Xujian. Document Layout Analysis Method Based on Edge Information Fusion and Multimodal Collaboration[D]. Hangzhou: Hangzhou Dianzi University, 2025.
[4] Tan Tian. Research on the High-quality Development of Ideological and Political Education in Colleges and Universities Empowered by Digitalization[D]. Guiyang: Guizhou Normal University, 2025.
[5] Xia Qiming, Liu Xiaoqiang. Design and Implementation of an AI Algorithm Visualization-assisted Teaching System[J]. Computer Knowledge and Technology, 2020(14): 41-43, 46.