

From tool to thinking: ComfyUI-based pathway for shadow puppetry style generation and reflections on practice

Xuefang Zhang¹, and Yiran Cheng^{2,*}

¹School of Journalism and Communication, Hangzhou City University, China

²Digital Humanities, University of Cambridge, Cambridge, United Kingdom

Abstract. To address the challenges of detail loss, stylistic generalization, and limited controllability in the digital generation of Chinese intangible cultural heritage “shadow puppetry”, this study proposes a generative pipeline developed through self-directed inquiry-based learning. The pipeline employs Flux.1 as the base model and integrates Low-Rank Adaptation (LoRA) fine-tuning within a ComfyUI-based workflow. Experimental results demonstrate that the proposed approach enables high-fidelity and controllable generation of shadow puppetry styles. Furthermore, an analysis of the learning process indicates that systematic tool application and reflective technical journals contribute to the development of learners’ computational thinking, cognitive transformation, and complex problem-solving abilities. This research not only presents a feasible technical solution for the digital preservation and creative reinterpretation of specific intangible cultural heritages but also provides an operable practical paradigm for empowering digital humanities education through AI-generated content (AIGC) technologies.

1 Introduction

As an intangible cultural heritage of humanity, Chinese shadow puppetry is distinguished by its unique aesthetic features, including light-and-shadow translucency, lateral silhouette composition, and intricate ornamental patterns. In the context of digital preservation and creative transformation, how generative artificial intelligence can be leveraged to achieve high-fidelity stylistic translation of shadow puppetry has remained a central challenge at the intersection of digital humanities and computer vision. While conventional diffusion models have demonstrated strong performance in general-purpose image generation, they often struggle with the high-contrast contours, delicate patterns, and dramatic light–shadow relationships characteristic of shadow puppetry, resulting in structural distortions, homogenized details, or insufficient stylistic specificity. Released in 2024, the Flux.1 model represents a significant advance in this regard. Owing to its large parameter scale and advanced architectural design, Flux.1 exhibits enhanced capabilities in interpreting complex visual semantics and maintaining output stability, thereby providing a

* Corresponding author: Yc677@cam.ac.uk

new technical foundation for fine-grained and controllable style generation in the digital innovation of intangible cultural heritage [1].

However, the introduction of advanced generative models does not automatically lower the threshold for practical application. On the contrary, the effective deployment of their technical potential often depends on the construction of highly flexible and programmable workflows. The shift from the graphically encapsulated interaction of Stable Diffusion WebUI to the node-based visual programming interface of ComfyUI represents, in essence, a paradigmatic transition from tool usage to logic construction. While this transition substantially enhances the controllability and extensibility of workflows—particularly for fine-tuning and stabilising specific styles through lightweight adaptation techniques such as LoRA—it also markedly increases learners’ intrinsic cognitive load. The logic of node connections, the transmission of parameters, and the visual management of the generative process require learners not only to understand model principles, but also to develop structured thinking and systematic debugging capabilities. Consequently, within the context of AIGC-enabled artistic creation, a profound tension emerges between the choice of technical pathways and learners’ cognitive adaptability.

Against this backdrop, this study adopts an interdisciplinary perspective bridging computer applications and educational technology, arguing that technical practice and learning processes should be organically aligned. On the one hand, through a self-directed exploratory learning trajectory, we examine both the effectiveness and limitations of combining Flux1 with LoRA for reproducing the stylistic characteristics of shadow puppetry, while offering an in-depth analysis of ComfyUI’s structural advantages in constructing complex and reusable generative logics. On the other hand, the technical practice itself is conceptualised as a form of instructional scaffolding, prompting reflection on its educational value in cultivating learners’ computational thinking, systematic problem-solving abilities, and autonomous learning strategies for engaging with complex tools. Rather than focusing solely on the technical question of how to implement, this research also foregrounds the cognitive process of how to learn, with the aim of proposing a transferable pathway for integrating advanced technical competencies with self-directed inquiry-based learning in the era of AIGC.

In summary, through a self-directed exploratory learning approach, this study aims to achieve the following core objectives:

To evaluate the visual fidelity, controllability, and creative potential of the Flux1 + LoRA technical pathway in stylised generation of shadow puppetry;

To deconstruct the design advantages of ComfyUI’s node-based workflows in supporting complex generative logic, iterative optimisation, and the encapsulation of procedural knowledge;

To analyse, from the perspective of educational technology, the cognitive barriers faced by novice learners when engaging with advanced AIGC tools, and to reflect on how this technical practice informs the cultivation of autonomous learning capabilities and cross-domain computational thinking, thereby providing a case-based reference for technology-enabled educational innovation.

2 Technical Framework and Theoretical Foundations

This study is grounded in an integrated technical system and theoretical logic that takes the Flux1 generative model as its foundational base, applies LoRA for low-parameter style fine-tuning, and employs ComfyUI to construct modular and interpretable generative workflows. This technical pathway is not only designed to enable high-quality digital generation of shadow puppetry aesthetics, but also—through its intrinsic structural

characteristics-serves as an exemplary model for technology-enabled, self-directed exploratory learning.

2.1 Flux.1: A Generative Foundation Based on Transformers and Flow Matching

The generative tasks in this research are built upon Flux.1 as the core foundation model. Unlike traditional Latent Diffusion Models (LDMs), which primarily rely on U-Net architectures for iterative denoising, Flux.1 introduces architectural and training paradigm innovations. It adopts a Transformer-based Diffusion Transformer (DiT) as its backbone, enabling bidirectional parallel attention computation over both image and text tokens. This design endows the model with superior semantic alignment capabilities when interpreting complex, non-photorealistic descriptions such as the semi-translucent texture and lateral silhouette characteristic of shadow puppetry.

In terms of training paradigm, Flux.1 incorporates flow matching, a technique aimed at learning a more direct linear interpolation path between the data distribution and the noise distribution. Compared with conventional diffusion processes grounded in stochastic differential equations, this mechanism significantly reduces the number of inference steps required for high-quality generation, while more effectively preserving high-frequency image details under accelerated sampling. Such properties are crucial for accurately reproducing the sharpness of intricate ornamental linework and the dramatic light-shadow contrasts intrinsic to shadow puppetry, thereby providing a robust and highly responsive foundation model for subsequent style-oriented fine-tuning.

2.2 LoRA: Parameter-Efficient Fine-Tuning for Stylised Generation

To precisely steer Flux.1's general generative capabilities towards the specific aesthetic domain of Chinese shadow puppetry, this study adopts Low-Rank Adaptation (LoRA) as a parameter-efficient fine-tuning strategy. The core mechanism of LoRA consists in freezing all pretrained model parameters and injecting trainable low-rank decomposition matrices into selected layers of the network, particularly the Transformer attention modules [2-3]. Formally, for a given Transformer layer, the hidden representation can be expressed as $h = W_0 x + \Delta W x = W_0 x + BA x$, where W_0 denotes the frozen pretrained weight matrix, x represents the input embedding, and $\Delta W = BA$ is the low-rank update composed of two trainable matrices A and B , with $\text{rank } R \ll \min(\text{dim}(W_0))$. By constraining model adaptation to this low-dimensional subspace, LoRA enables the targeted injection of stylistic features-such as the ornamental motifs, line qualities, and silhouette characteristics of shadow puppetry. Within an educational technology context, this method offers a dual advantage. First, in terms of technical performance, LoRA requires the training of only a minimal number of parameters to achieve targeted style feature injection using small-scale yet high-quality domain-specific datasets. This effectively mitigates risks of overfitting and catastrophic forgetting, ensuring that while the model acquires the distinctive patterning and chromatic characteristics of shadow puppetry, it retains its original compositional and semantic understanding. Second, from a learning-process perspective, the relatively low computational demands and data annotation costs of LoRA fine-tuning make it feasible to conduct model customisation and experimental inquiry focused on specific artistic styles within constrained learning environments and timeframes. This aligns closely with the feasibility and iterative experimentation requirements inherent to self-directed exploratory learning.

2.3 ComfyUI: An Implementation Environment as a Visual Logic Scaffold

To achieve fine-grained control and a deeper understanding of the Flux.1+ LoRA generative pipeline, this study adopts ComfyUI as its primary implementation and experimental environment. At its core, ComfyUI is a node-based visual dataflow programming interface that decomposes the complete image generation process into discrete yet connectable modules (nodes) [4], such as model loading, prompt encoding, latent-space sampling, and image decoding.

This design constitutes a crucial form of technical–cognitive scaffolding within the present research. First, by visualising the internal dataflow of AIGC systems—such as transformations between latent space and pixel space—ComfyUI compels learners to engage with the inputs, outputs, and purposes of each procedural step, thereby concretising the otherwise abstract notion of “generation” into an observable and debuggable chain of logic. Second, its highly modular architecture supports flexible workflow orchestration, facilitating the integration of LoRA adapters, ControlNet, and other control networks to enable multi-dimensional constraints and iterative optimisation over both stylistic and structural aspects of the generated outputs. Finally, each node connection and parameter adjustment instantiates a clearly articulated hypothesis–verification cycle, in which errors (e.g. data type mismatches) are immediately surfaced. This strongly supports debugging-oriented inquiry-based learning, effectively managing extraneous cognitive load during technical practice while directing learners’ attention towards logic construction rather than interface manipulation.

2.4 Integrated Framework: Aligning Technical Pathways with Learning Trajectories

In summary, this study establishes a three-layer technical framework comprising a powerful foundation model (Flux.1), precise style adaptation (LoRA), and controllable logical implementation (ComfyUI). From a technical standpoint, this framework systematically addresses key challenges in shadow-puppetry stylised generation, including semantic alignment, detail fidelity, and controllable synthesis.

Viewed through the lens of educational technology, however, this framework also delineates a coherent pathway for self-directed learning. It guides learners from an initial understanding of state-of-the-art generative model principles, through the mastery of efficient fine-tuning methods tailored to specific domains, and ultimately towards advanced engagement with complex systems via visual programming, enabling the logical decomposition and reconstruction of generative processes. In doing so, the framework transforms high-dimensional technical practice into a series of manageable, investigable, and reflexive learning tasks. As a result, the acquisition of technical tools and the cultivation of computational thinking are synchronously realised through the practical resolution of authentic and complex problems in the digitalisation of cultural heritage.

3 Experimental Design and Results Analysis

3.1 Experimental Environment

The experimental platform employed in this study is the Zhiling Trainer, which supports multiple mainstream LoRA training scenarios. Its integrated training pipeline significantly reduces the complexity of environment configuration, enabling researchers and creators to rapidly conduct customised model training. This characteristic aligns well with educational

technology principles that emphasise lowering initial technical barriers and allowing learners to focus on core exploratory processes. Based on the requirements of the present study, the hardware configuration consisted of an NVIDIA GeForce RTX 4090 GPU with 24 GB of VRAM and 64 GB of system memory.

The training data comprised a self-constructed, small-scale yet high-quality dataset, including 15 image–text pairs of Chinese shadow puppetry characters representing different role types such as sheng, dan, jing, and chou. All images were uniformly processed to a resolution of 1024×1024 pixels. The key training parameters were set as follows: a total of 5,760 training steps, 24 epochs, a batch size of 1, the AdamW 8-bit optimiser, and a learning rate of 0.0001. This small-sample, multi-epoch training strategy was designed to encourage the LoRA model to internalise the essential stylistic characteristics of shadow puppetry, rather than merely memorising the training samples.

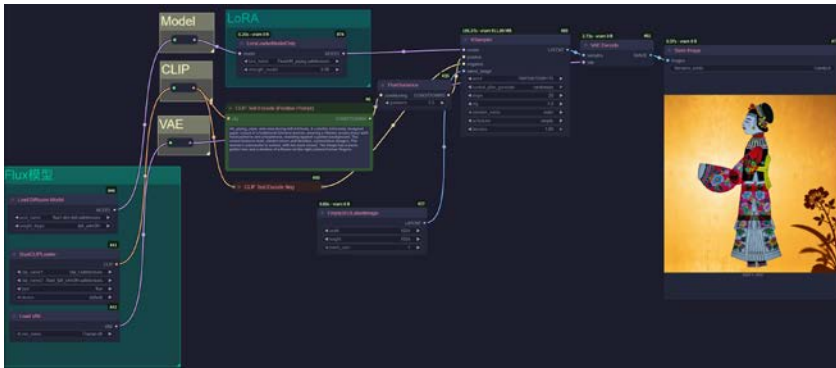


Fig. 1. A modular text-to-image workflow constructed in ComfyUI.

To evaluate the trained LoRA model, a modular text-to-image workflow was constructed in ComfyUI (see Figure 1). This workflow visualises and logically decouples the full pipeline of base model loading (Flux.1) → style injection (LoRA) → conditional encoding (CLIP) → sampling and generation (KSampler) → decoding and output (VAE). Using the general-purpose Flux.1 model as the generative foundation, the workflow introduces shadow-puppetry style through lightweight, targeted LoRA adaptation, while textual encoding provides semantic constraints that guide the diffusion sampling process to synthesise desired features within latent space, which are then decoded into visual outputs. Beyond serving as a tool for producing generative results, this environment functions as a cognitive scaffold for understanding internal data flows and module interactions within AIGC systems, rendering the otherwise abstract generative process observable, debuggable, and optimisable. To establish a clear and verifiable technical pathway, this study designed a four-stage experimental workflow (see Figure 2).

3.2 Results Analysis

The experimental results indicate that the Flux1 model augmented with the shadow-puppetry LoRA demonstrates a high level of artistic fidelity across multiple dimensions.

Texture and linework: The model successfully reproduces the characteristic contours of shadow-puppet figures and the traditional floral motifs on costumes. The generated linework is sharp and well-defined, with no noticeable blurring or unintended merging of details.

Colour and lighting: The lighting effects typical of shadow-puppetry backdrops are effectively simulated, with warm colour palettes and high-contrast light–shadow relationships that align closely with the established aesthetic conventions of the art form.

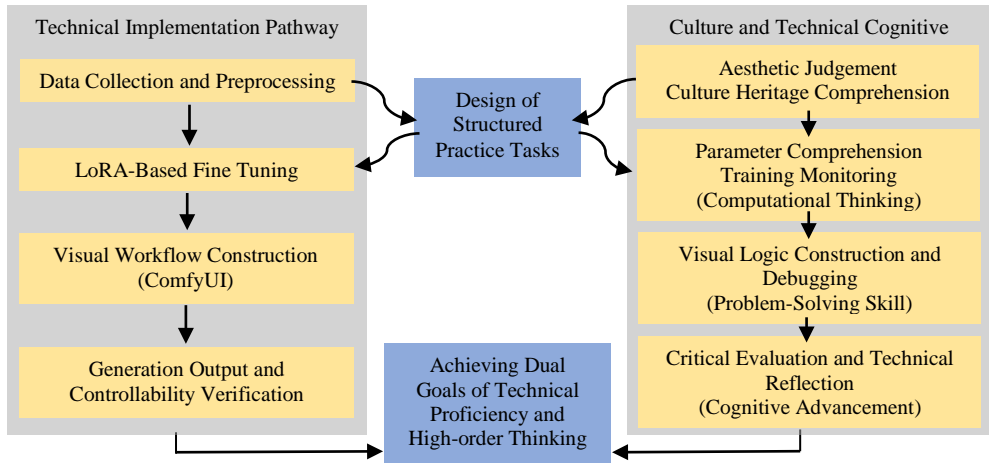


Fig. 2. Integrated technical-cognitive process workflow.

Structural stability: Full-body character generations exhibit well-proportioned limbs and natural joint connections, overcoming the structural distortions at articulation points that are commonly observed in conventional models. This demonstrates that the Flux.1 + LoRA combination is capable of preserving morphological plausibility while maintaining strong stylistic coherence.

To further examine the degree of control afforded by the LoRA model during generation, the experiment systematically adjusted the LoRA blending weight to values of 0 to 1.0, in order to observe the evolution of generated outputs. As illustrated in Figure 3, variations in blending weight exert a pronounced influence on both stylistic fidelity and creative flexibility.



Fig. 3. Generation results of the LoRA model at different weight settings.

High-weight range (1.0-0.8): Generated images closely replicate the stylistic features of the training data in terms of facial characteristics, costume forms and patterns, and bodily postures, reflecting a strong degree of stylistic constraint.

Mid-weight range (0.7-0.4): The primary costume silhouettes and dominant colour schemes are retained, while discernible morphological variations emerge in ornamental patterns and fine details.

Low-weight range (0.3-0.0): Stylistic features are substantially attenuated, persisting only in partial colour accents or contour cues. The overall imagery increasingly resembles the general-purpose outputs of the base model, with a marked tendency towards freer creative generation.

In summary, this study validates the technical feasibility of the proposed framework for shadow-puppetry style generation. At the same time, the experimental design itself exemplifies a pedagogical approach that translates complex technical systems into structured and investigable tasks, thereby offering both empirical evidence and methodological reflection for the application of AIGC technologies in digital humanities education.

4 Reflection on Technical Practice: From Tool Usage to Systemic Understanding

Based on the complete experimental process outlined above, this study documents the technical challenges and learning trajectories encountered by learners during the construction of a complex AIGC toolchain. The following section synthesises key insights drawn from systematic technical reflection notes.

4.1 Accessibility Challenges of the Technical Ecosystem and Cognitive Reconfiguration

In the initial phase of the experiment, ComfyUI was deployed in a non-standard environment (macOS). When attempting to run the LoRA-Training extension within the ComfyUI environment, a series of challenges emerged, including path resolution errors, compilation failures of dependency libraries (such as wandb), and severely reduced training efficiency due to the lack of CUDA support, which forced CPU-based computation. These issues extended well beyond routine software operation, compelling the researchers to engage directly with low-level system configurations through a form of “reverse construction”, involving manual creation of symbolic links, environment variable configuration, and installation of underlying compilers and toolchains (e.g. the Go programming environment).

These difficulties expose a structural bias embedded within many contemporary open-source AIGC toolchains: their design is typically optimised around Windows/Linux systems and the NVIDIA GPU ecosystem. Such hardware-centricism effectively constitutes an implicit barrier to technical access, imposing significantly uneven system friction and time costs on users operating across different hardware platforms. As a result, the often-invoked notion of “technological democratisation” is, in practice, accompanied by a form of infrastructure-based exclusivity.

At the same time, it is important to note that the process of overcoming these obstacles—while substantially increasing initial cognitive load—strongly motivated learners to move beyond graphical interfaces and engage with the underlying operational logic of the tools, their module dependency structures, and principles of cross-platform compatibility. In this sense, the technical friction functioned as a catalyst for deeper cognitive reconfiguration,

enabling learners to develop a more systemic and critical understanding of the AIGC technical ecosystem.

4.2 Debugging as a “Desirable Difficulty” for Cognitive Construction

ComfyUI’s node-based workflow decomposes the generative process into discrete, visualised data streams. While this decomposition increases operational complexity, it simultaneously creates a form of desirable difficulty. When node connections fail or data transmission breaks down, the system’s explicit error messages point directly to the locus of logical discontinuity. As a result, debugging ceases to be a process of blind trial and error and instead becomes a targeted activity of principle-based analysis. Learners are compelled to clarify the substantive meanings and functional roles of abstract concepts—such as latent space, conditioning vectors, and sampling steps—within the dataflow, thereby transforming black-box operations into white-box logic.

This process closely aligns with constructivist learning theory, which posits that knowledge is not passively received but actively constructed through engagement with authentic and complex problem-solving activities. Each adjustment of pathways and repair of dependencies constitutes a dialogue between the creator and the technical system, through which learners transition from passive “tool users” to active investigators capable of understanding and reconstructing technical structures. Each successful debugging episode deepens comprehension of the operational mechanisms underlying the AIGC pipeline, gradually assembling a more robust and transferable map of computational thinking.

4.3 The Dual Role of Open-Source Communities: Scaffolding and Bias

Self-directed learning relies heavily on external resources. In this practice, custom node implementations on GitHub and LoRA models available on platforms such as Civitai and Liblib functioned as indispensable learning scaffolds. By lowering the threshold for creation from scratch, these resources enabled rapid exploration through reuse and comparative experimentation.

At the same time, cross-platform comparative analysis revealed that models are not merely products of mathematical optimisation but also encode the cultural assumptions embedded in their training data. In this sense, open-source communities themselves function as carriers of cultural values. For example, platforms such as Civitai, dominated largely by Western communities, and Liblib, which exhibits a higher degree of localisation, display marked differences in training data sources, aesthetic preferences, and cultural orientations. Models hosted on Civitai tend to reflect Western aesthetic conventions and linguistic corpora, whereas Liblib aggregates a substantial number of models characterised by Eastern cultural features. These differences extend beyond surface-level visual styles or character traits, revealing deeper divergences in the cultural and semantic orientations underlying AI image generation—what may be termed an aesthetic ideology encoded at the data level. In other words, variations in data distributions and model biases across platforms effectively produce a form of cultural–geographical stratification within the global AI art ecosystem.

Consequently, the use of open-source models cannot be regarded as a culturally neutral technical act; rather, it entails a process of cultural adoption and negotiation. This insight underscores the need for educators and technical practitioners to adopt a critical perspective when drawing upon community resources for teaching or creative practice, recognising the socio-cultural constructions embedded within technical artefacts.

Taken together, the technical practice documented in this study constitutes an intensive form of cognitive training. It demonstrates that mastery of advanced AIGC technologies hinges not merely on operational proficiency, but on the capacity to transform challenges encountered in tool usage, environment configuration, and resource selection into opportunities for deepened systemic understanding and the cultivation of critical thinking. The role of educators, in this context, should be to design structured learning pathways that guide learners through such productive complexity, converting cognitive load into cognitive momentum, and ultimately fostering hybrid practitioners who are not only adept at deploying tools but also capable of interrogating their technical logics and cultural implications.

5 Conclusions and Implications

5.1 Conclusions

Through the concrete practice of shadow-puppetry stylised generation based on Flux1 and ComfyUI, this study systematically validates the feasibility of a technical pathway that integrates an advanced foundation model, a parameter-efficient fine-tuning method (LoRA) as a style adaptation mechanism, and a node-based visual programming environment (ComfyUI) as a framework for logical implementation. The experimental results demonstrate that this approach is capable of generating shadow-puppetry-style images with a high degree of artistic fidelity in terms of textural detail, colour and lighting, and structural stability, thereby successfully achieving the digital stylistic translation of a specific form of intangible cultural heritage. More importantly, the research consistently integrates perspectives from both computer applications and learning technologies, treating the technical practice itself as an object of inquiry and reflection. In doing so, the study not only fulfils its empirical technical objectives but also yields methodological and pedagogical insights that extend beyond the use of any single tool.

At its core, this research represents an interdisciplinary endeavour that fuses technical validation with cognitive inquiry. The proposed stylised generation pathway offers a practical solution for addressing non-photorealistic heritage styles characterised by high levels of detail and distinctive light–shadow aesthetics, such as shadow puppetry. At the same time, its modular and interpretable workflow provides a reusable methodological reference for the application of AIGC technologies in other vertically specialised artistic domains. Furthermore, the study illustrates how the “desirable difficulties” encountered in complex technical environments can naturally foster constructivist learning, prompting learners to transition from tool users to logic builders. It also demonstrates how data and cultural biases embedded within open-source model ecosystems can serve as authentic contexts for cultivating learners’ critical technical literacy.

5.2 Implications for Art and Design Education in the Intelligent Era

Drawing on the findings and reflective insights of this study, the following implications are proposed for contemporary education in digital art, intelligent media, and related fields.

5.2.1 Reconstructing Educational Objectives: From Software Operation Skills to Computational Thinking and Systemic Understanding

Art and design education should move beyond training students in the operational

mechanics of specific software interfaces, and instead prioritise the cultivation of core capacities for understanding, constructing, and controlling digital creative systems. Tools such as ComfyUI can be employed to visualise and explicitly teach foundational concepts—including latent space, conditional guidance, and iterative sampling—thereby elevating procedural knowledge of how to generate into principled understanding of why generation operates in this manner. Such a shift is essential for addressing rapid technological evolution and for fostering learners’ sustainable creative capacities.

5.2.2 Innovating Pedagogical Approaches: Deepening Project-Based Learning (PBL) with Cultural Purpose and Technical Critique

The design of integrative projects centred on the digital translation of intangible cultural heritage can situate technical learning within authentic, complex, and culturally meaningful tasks. This approach not only enhances intrinsic learner motivation, but also enables the organic integration of multidisciplinary knowledge and skills through problem-solving. At the same time, such projects should explicitly encourage critical examination of data biases, cultural representations, and structural dynamics embedded within technological systems, thereby nurturing responsible and critically informed innovation practices.

5.2.3 Empowering Learning Processes: Integrating “Learning by Doing” with “Learning by Reflecting” and Valuing Process-Oriented Reflection

Learners should be encouraged to document debugging, error resolution, and optimisation processes in the form of technical reflection notes. Beyond serving as records of problem-solving, these notes function as exercises in metacognitive development, helping learners consolidate fragmented experiences into structured knowledge and cultivate the ability to monitor and adapt their own learning strategies. In this context, educators should assume the role of cognitive coaches, guiding learners to extract transferable principles from moments of technical frustration and to complete the cognitive transition from experience to insight.

References

1. W. Ye, T. Zheng, Y. Qi, W. Zhao, X. Wang, X. Zhao, J. He, Y. Zheng, D. Wang, arXiv:2505.23831 (2025)
2. E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Proc. Int. Conf. Learn. Represent. (2022)
3. X. Zhang, Heritage 7, 11 (2024)
4. S. Ali, et al., Proc. AAAI Conf. Artif. Intell. 38, 21 (2024)
5. R. Knabäck, Impact of Gaussian Noise on LoRA Training for FLUX.1 (2025)
6. A. Hingle, Companion Proc. ACM Int. Conf. Supporting Group Work (2025)
7. L. Chen, P. Chen, Z. Lin, IEEE Access 8, (2020)
8. C. Zheng, et al., Proc. CHI Conf. Hum. Factors Comput. Syst. (2024)