Optimising ancient costume image generation using the stable diffusion model: a focus on dynastic characteristics

DOI: 10.35530/IT.077.03.202570

JIAN HUA
ZHI LI
RUIHONG CHEN
YU CHEN

ABSTRACT – REZUMAT

Optimising ancient costume image generation using the stable diffusion model: a focus on dynastic characteristics

The source of the generated images of ancient costumes is misplaced because the process of generating ancient costume effect images cannot accurately capture the distinctive features of different dynasties. By leveraging the Stable Diffusion model, this study organises costume characteristics across different dynasties into 163 textual prompts, drawing on historical literature and classical scrolls. By matching these prompts with image feature vectors, a new token embedding layer V is introduced, which is optimised together with the cross-attention layer parameters W^k and W^v. Then, the model was fine-tuned using the Low-Rank Adaptation (LoRA) model to reduce training costs while maintaining historical fidelity. The results demonstrate that the optimised model can generate costume images that align with the corresponding dynastic and ethnic characteristics based on textual prompts. Validation experiments across the Tang, Song, and Ming dynasties show that the model achieves relatively low Kernel Inception Distance (KID) and Maximum Mean Discrepancy (MMD) values, indicating its effectiveness in generating ancient costume images. This study not only optimises the generation of ancient costume effect images but also holds reference value for the digital preservation and protection of costume cultures from other dynasties and ethnic regions.*

Keywords: *ancient costumes, image generation, intelligent design, stable diffusion model, text-to-image*

Optimizarea generării imaginilor cu costume antice folosind modelul Stable Diffusion: un accent pe caracteristicile dinastice

Sursa imaginilor cu costume antice generate este eronată din cauza incapacității de a surprinde cu precizie trăsăturile distinctive ale diferitelor dinastii în timpul procesului de generare a imaginilor cu efecte de costume antice. Folosind modelul Stable Diffusion, acest studiu organizează caracteristicile costumelor din diferite dinastii în 163 de prompturi textuale, pe baza analizei literaturii istorice și a pergamentelor clasice. Prin potrivirea acestor prompturi cu vectorii de caracteristici ai imaginilor, este introdus un nou strat de încorporare a token-urilor V, care este optimizat împreună cu parametrii stratului de atenție încrucișată Wk și Wv. Apoi, modelul a fost ajustat folosind modelul Low-Rank Adaptation (LoRA) pentru a reduce costurile de antrenare, menținând în același timp fidelitatea istorică. Rezultatele demonstrează că modelul optimizat poate genera imagini cu costume care se aliniază cu caracteristicile dinastice și etnice corespunzătoare, pe baza prompturilor textuale. Experimentele de validare din dinastiile Tang, Song și Ming arată că modelul atinge valori relativ scăzute ale Kernel Inception Distance (KID) și Maximum Mean Discrepancy (MMD), indicând eficacitatea sa în generarea de imagini cu costume antice. Acest studiu nu numai că optimizează generarea de imagini cu efecte de costume antice, ci are și o valoare de referință pentru conservarea digitală și protecția culturilor vestimentare din alte dinastii și regiuni etnice.*

Cuvinte-cheie*: costume antice, generarea de imagini, proiectare inteligentă, modelul Stable Difussion, text-imagine*

INTRODUCTION

As an important part of the world's cultural heritage, costumes contain historical narratives and cultural significance. In the digital era, the digital reproduction of ancient costumes has become an essential strategy for preserving and sharing traditional history and culture [1]. The text-to-image model generates semantically compliant image content based on textual prompts, achieving semantic mapping across different modalities. This new method has become a powerful means of generating images of costumes [2–4]. However, it is difficult for existing models to capture the complex details and temporal nuances of ancient costumes, especially in representing dynasty-specific design elements, fabric textures, and chronological changes in costume construction, which limits their effectiveness in cultural preservation.

Although models such as Generative Adversarial Networks (GANs) [5] and CLIP [6] have demonstrated exceptional capabilities in general image synthesis, their application in generating specific ancient costumes is limited. For example, Patashnik et al. [7] proposed a joint embedding method based on GAN and CLIP for text-driven image generation,

constraining the matching of geometric relations between the text vector difference (Δt) and the image vector difference (Δs). This method performs well regarding global semantic alignment but still suffers from distortion in generating high-frequency details. ClothGAN [3] improved the generation of Dunhuang mural costumes but requires more training data than the Stable Diffusion (SD) model [8]. A similar problem was addressed by Avrahami et al. [9], who segmented image regions based on localised textual descriptions, thus enabling finer spatial control. While this approach improves spatial coherence, it can miss or incorrectly infer key historical features of costumes. In addition, Saito et al. [10] proposed a Compressed Image Retrieval (CIR) model that improves text-image matching without labelling training data, but sacrifices computational efficiency and restricts scalability and accessibility.

In recent years, advances in generative artificial intelligence have had a significant impact on various fields, including textile digitisation. In particular, it has been applied in the preservation and digital reconstruction of cultural heritage [1]. Liu et al. [11] used deep learning and virtual fitting to reconstruct costumes from the Five Dynasties period, revealing historical fashion trends. The image generation model was applied to the restoration and design of specific costumes [12], contributing to the authenticity of digital collections. Zhuo Shi et al. [13] refined diffusion models to generate accurate depictions of traditional Yao costumes. In contrast to traditional generative models, such as GANs and VQ-VAE [14], SD has emerged as a dominant framework for text-to-image generation, particularly in terms of its denoising-based stability [8] compared to GANs [15]. One of its main advantages lies in its conditional generation and noise-guided optimisation, which enable precise control over image features based on specific textual prompts and enhance the fine details and fidelity of the generated images [16]. SD's noise-guided optimisation enables finer control over historical details, such as textures and colour patterns, making it a promising candidate for ancient costume generation. Nevertheless, current models often struggle to capture the complex and context-specific features of ancient costume, particularly those tied to regional or temporal variations. While the integration of textual conditional control embedding layers has shown good potential for application, SD's reliance on pre-trained generators and the CLIP algorithm limits its ability to adequately capture the complex and diverse characteristics inherent in ancient costumes [17, 18].

As a consequence, this paper proposes two synergistic innovations:

(1) A V* token embedding layer that dynamically maps dynasty-specific textual prompts to latent space markers for more accurate generation of era-identifying features.
(2) A parameter-efficient fine-tuning framework based on the Low-Rank Adaptation (LoRA), which reduces computational requirements while maintaining the performance of multi-dynasty costume generation.

MODEL FOR COSTUME IMAGE GENERATION

This section outlines the proposed approach, which consists of several key steps: analysis and extraction of prompts, optimisation of feature parameters, and image generation model improvement.

Dataset construction and textual prompts

The formulation of textual prompts is key to guiding text-to-image models in generating contextually accurate costume images. The framework employs a dual-prompt composition that utilises positive prompts to specify desired attributes such as colour, style, material, and accessories to ensure historical accuracy. Negative prompts are also used to filter out inappropriate elements, such as modern zippers and synthetic fabrics, to eliminate inconsistencies that greatly affect the model's ability to generate visually and contextually appropriate images.

Aiming at the issues of multi-dynasty costume style changes and data scarcity in ancient costume research, this study constructs a multi-dynasty figure costume image dataset. The dataset systematically integrates 2031 image samples (each image contains more than five attributes, such as collar shape and fabric texture) from digital archives of museum collections, illustrations of historical literature, and paintings of ancient figures, covering the three key periods of costume changes in typical Chinese traditional culture, namely the Tang, Song, and Ming. The specific dataset processing is described in the section below. For example, the Song Dynasty dataset (figure 1) includes a round-collared robe, a spread-feet scarf, and the holding of a ceremonial tablet, which are key to the historical accuracy of costumes. In addition, we compiled 163 textual prompts from historical scrolls and literature, covering the key features of these dynasties' costumes (table 1). This specialised dataset, along with detailed prompts focusing on collar styles, fabric types, and accessory

Fig. 1. Representative training images of Song Dynasty men's costumes showing key attributes such as round-collared robes, ceremonial tablets, and period-specific footwear

Table 1

DESCRIPTION OF PROMPTS FOR CHINESE SONG DYNASTY COSTUMES (MALE)

Category	Examples	Key features
Head ornament	scarf	spread feet scarf
		stiff feet scarf
	kerchief	kerchief
	chignon	chignon
	coronet	coronet
Under-garment	belly lock	belly lock
	undergarment	undergarment
Pant	pleated skirt	pleated skirt
Footwear	gleditsia boots	gleditsia boots
	bow shoes	bow shoes
	toe shoes	toe shoes
	square-toe shoes	square-toe shoes

details, was incorporated into the model training process to improve its ability to generate accurate and contextually appropriate descriptions of ancient costumes.

Token embedding (V*) for era-discriminative control

In the diffusion model, the cross-attention layer is important for ensuring that textual prompts are consistent with visual features [8]. Nupur et al. [19] studied the tuning of model parameters for introducing new features and observed significant variation in the parameter change rates across layers of the U-Net [20] in the SD model. These parameters were derived from three types of training layers: the cross-attention layer (for interaction between images and text), the self-attention layer (for correlation within images), and the remaining parameters (convolutional blocks and normalisation layer in the LDM model U-Net). Analysis of the rate of change in parameter weights of the loss function across different layers $\Delta t = \|\theta_t' - \theta_t\| / \|\theta_t\|$ (where θ_t' and θ_t represent model parameters for the first layer update and pre-training, respectively) reveals that the cross-attention layer undergoes substantial Δ change during model tuning, despite constituting only 5% of the total parameters. This observation motivated us to focus on optimising these layers for era-discriminative control.

Therefore, in this study, we adjusted the input costume text feature condition to modify the image features of the U-network based on the cross-attention layer $Attention(Q,K,V)$ [21]. Given the text $c \in \mathbb{R}^{s\times d}$ and image $f \in \mathbb{R}^{(h\times w)\times l}$ features, the single-head cross-attention operation involves $Q = W^q f$, $K = W^k c$, $V = W^v c$ and a weighted sum of value features.

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d'}}\right)\cdot V \qquad (1)$$

where W^q, W^k, and W^v map inputs to the query, key, and value features, respectively, and d' is the output dimension of the key and query features. The purpose of tuning is to update the mapping from custom text features to the distribution of custom-image features, which are only input into the key-values (KV) layer of the cross-attention layer, and thus optimised by injecting new costume text features along with the W^k and W^v parameters of the cross-attention layer in the SD model when tuning the model in this study. This procedure effectively updates the model to incorporate new costume text-image pairs.

To address the problem of temporal feature obfuscation, this study introduces the V^* token embedding. This is a trainable layer inspired by Abdal et al. [22] that injects dynasty-specific semantic tokens into the cross-attention mechanism. V^* was optimised as a textual conditional embedding layer and integrated with the relevant parameters of the cross-attention layers W^k and W^v. The implementation of V^* involves three key steps:

Token initialisation

A single collected costume image contains multiple new costume features in different categories. Therefore, the joint training of each costume text feature is required to tune the model. Different token symbols Vi^* are used during token initialisation to represent different costume text features. The token V^* is initiated using the mean and standard deviation of the CLIP text embeddings extracted from historical references to ensure their semantic relevance to the target dynasty.

$$V^*_{init} = \mu_{CLIP} + \sigma_{CLIP} \odot \epsilon,\ \epsilon \sim N(0,1) \qquad (2)$$

where μ_{CLIP} denotes the mean value of the CLIP text embedding of the costume description for a specific historical period. σ_{CLIP} is the standard deviation of the CLIP embedding, reflecting the degree of feature discretisation of the textual description.

Feature augmentation

Given the text $c \in \mathbb{R}^{s\times d}$, and image feature $f \in \mathbb{R}^{(h\times w)\times l}$, the cross-attention operation is enhanced by connecting V^* with c.

$$K = W^k \cdot [c \oplus V^*],\ V = W^v \cdot [c \oplus V^*] \qquad (3)$$

where $[c \oplus V^*]$ denotes the channel connection. The tandem features are then processed by the key matrix (W^k) and value matrix (W^v) to form the key (K) and value (V) vectors of the attention mechanism.

Layer-specific adaptation

For each cross-attention layer y, only the key and value matrices associated with the costume text features are updated to form a unified parameter set $\{W^k_{0,y}, W^v_{0,y}\}^L_{y=1}$. If n new costume features are added ($n \in \{1...n\}$), the corresponding update matrix is defined as $\{W^k_{n,y}, W^v_{n,y}\}^L_{y=1}$. The subscript y is omitted below and denoted by $W^{K,V}$ which only the mapped text features are modified. The training objective combines the denoising loss with a forgetting penalty.

$$\min_{W^{K,V}_t \in \theta} \mathcal{L}_{diff}(x,\theta) + \lambda \mathcal{L}_{forget}(W^{K,V}_{t-1}, W^{K,V}_t) \qquad (4)$$

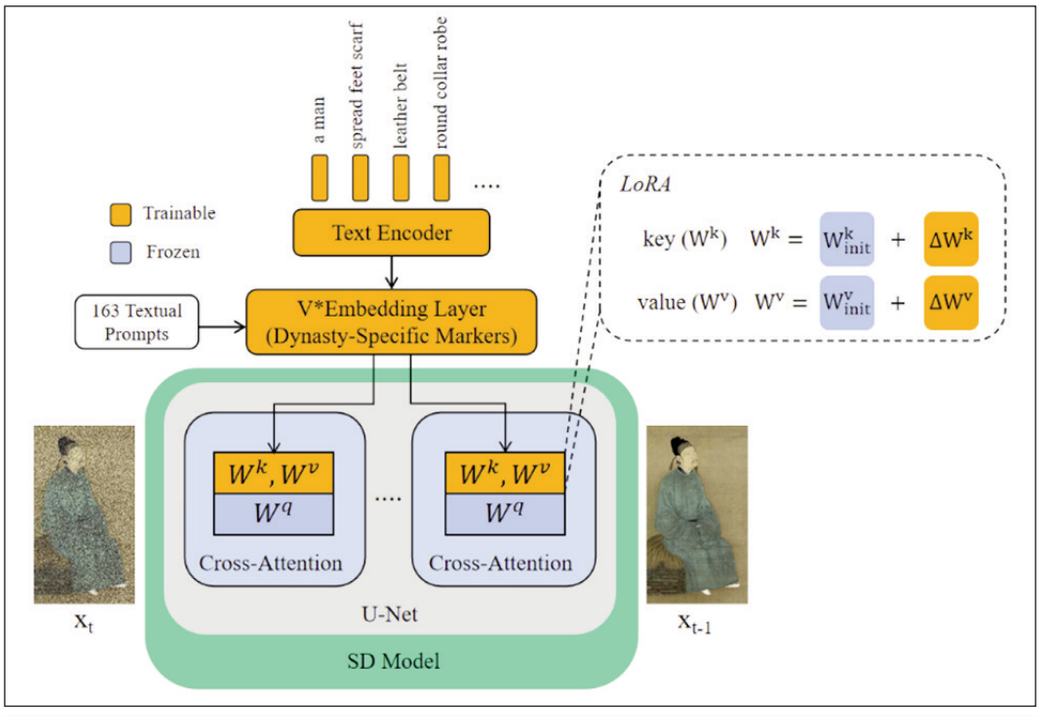

Fig. 2. Proposed model architecture with *V** embedding and LoRA module

where x is the new text feature data, and $\mathcal{L}_{diff}$ is the updated SD model loss function, which is the standard denoising objective for image generation. Model θ is not modified, and $\mathcal{L}_{forget} = \| W_t^{K,V} - W_{t-1}^{K,V} \|_2^2$ prevents disastrous forgetting by limiting the deviation from the pre-training weights. λ is the hyper-parameter chosen by exponential random search ($\lambda \in [0.1,1.0]$).

The proposed architecture integrates the CLIP text encoder and token embedding (*V**) to condition the cross-attention layers of U-Net, as illustrated in figure 2. The CLIP encoder first converts the input text prompts into semantic embeddings, which are connected to trainable *V** tokens to form enhanced text features. These features interact with the latent images through the cross-attention mechanism, where the key and value matrices are adaptively updated using the LoRA decomposition ($W = W_0 + BA$) to inject era-specific features.

Training proceeds in two phases:

1) Warm-up Phase: In the first five epochs, only the *V** embedding is trained with a forgetting penalty factor λ = 0.1 to ensure a stable initialisation while preserving the pre-training knowledge.
2) Joint Phase: 20 joint optimisations of *V**, W^k, and W^v using the AdamW-8bit optimiser with a learning rate of 5×10−4 and cosine decay scheduling. This V*-enhanced cross-attention was further optimised by LoRA, as detailed below.

Cross-attention optimization with LoRA

Full fine-tuning of large-scale diffusion models generated by ancient costumes incurs excessive computational costs. Therefore, we adopted LoRA [23], a parameter-efficient fine-tuning strategy that restricts weight updates to low-rank subspaces. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, the update decomposition is

$$W = W_0 + \Delta W = W_0 + BA$$
$$B \in \mathbb{R}^{d\times r},\ A \in \mathbb{R}^{r\times k},\ r \ll \min(d,k) \quad (5)$$

where ΔW is the parameter to be updated.

In the implementation mechanism, the original SD parameters (W_0) are frozen to preserve pre-trained knowledge and reduce catastrophic forgetting. During the training process, only low-rank matrices *A* and *B* are updated to achieve an efficient adaptation. The modified output for the input x is calculated as

$$h = W_0 x + \Delta W x = W_0 x + BAx \quad (6)$$

where $\Delta W = BA$ denotes the weight update, and h denotes the final feature embedding.

Subsequently, we applied LoRA to the key and value matrices of the cross-attention layer to strengthen era-specific feature alignment. The incremental update formula for layer t is

$$W_t^{K,V} = W_{init}^{K,V} + \sum_{t'=1}^{t-1} A_{t'}^{K,V} B_{t'}^{K,V} + A_t^{K,V} B_t^{K,V} \quad (7)$$

where $W_{init}^{K,V}$ denotes the pre-trained weights and $A_t^{K,V} \in \mathbb{R}^{D_1\times r}$, $B_t^{K,V} \in \mathbb{R}^{r\times D_2}$ are low-rank adapters.

Meanwhile, we imposed the L2 regularisation penalty $\| A_t^{K,V} B_t^{K,V} \|_F$ with a coefficient λ = 0.01 to mitigate overfitting on the limited training data.

The optimisation method of the LoRA-enhanced cross-attention mechanism dynamically integrates era-specific textual prompts and visual features, which can effectively adapt to ancient costume features and maintain the generalisation ability of the base model.

EXPERIMENTS AND RESULTS

To evaluate the performance of the model for generating costume renderings of ancient Chinese costumes, qualitative and quantitative evaluation analyses were conducted.

Training set preparation

We proposed a specialised dataset of 2,031 high-resolution images (512 × 768 pixels) divided into Tang (643 images), Song (744 images), and Ming (643 images) dynasties. Training set (1625 images), validation set (203 images), and test set (203 images).
We further used data augmentation approaches to increase the diversity and robustness of the dataset. First, we simulated the performance of costumes from different viewpoints and poses through geometric transformations, such as random cropping, rotation, and scaling. Second, the colour space was adjusted (brightness, contrast, and saturation) to improve the ability of our model to adapt to lighting changes. Finally, a slight noise was introduced into the image to improve the robustness of the model for low-quality input data. In addition, we introduce a regularisation term in Section 2.3 to prevent the model from over-relying on specific features. Although the data augmentation approach mitigates the limitations of the small dataset to some extent, we will still analyse the risk of overfitting in Section 4 and explore the importance of expanding the size of the dataset in future studies.

Experiment details

The experiments were conducted on an NVIDIA RTX 4090 GPU (24GB VRAM) using PyTorch 2.1.2 with CUDA 12.1 acceleration. Based on the community-validated chilloutmix_NiPrunedFp32Fix basic model, we implemented our adaptation framework using the Kohya-ss toolkit. Only 75MB of model weights are updated, which requires less memory for the model. In the specific model parameter tuning experiments, we determined the parameters dim = 32, alpha = 32, optimiser type (AdamW8bit), and a batch size of 1 as the optimal parameters for the model by comparing the clarity, the restoration of costume texture details, and the differentiation of dynastic features of the ancient costume images generated by the model. The number of repetition steps (repeat = 10), the number of training rounds (epoch = 20), the U-Net learning rate (5e–4), and the text learning rate (5e–5) were determined by comparing the loss function size of the model. These parameter settings ensure that the model achieves better results in learning costume features, and the generated images have the advantages of high clarity and obvious dynastic costume features.

Evaluation metrics

The evaluation protocol uses three key metrics: CLIP Alignment Scores for semantic consistency between generated images and text prompts, Kernel Inception Distance (KID) to quantify divergence from real images in the feature space, and Maximum Mean Discrepancy (MMD) to evaluate distributional similarity. Metric selection followed recent advances in text-to-image evaluation [24–26], with lower KID/MMD and higher CLIP scores indicating better performance. All metrics were calculated using five random seeds on a standardised test set to ensure statistical significance.

Fig. 3. Qualitative comparison of generated costumes across dynasties (From left to right: Real image, This model, Dreambooth, Textual Inversion, StyleCLIP)

Comparative experiments

In the comparative experiments, we chose to make a comprehensive comparison with three representative customised generation methods: Dreambooth [27], Textual Inversion [28], and StyleCLIP [7]. All models were based on the same training data, hardware configuration, and hyperparameter settings. Figure 3 shows the results of generating Tang, Song, and Ming dynasty costumes using real images as a reference. Despite Dreambooth's high parametric efficiency (32MB model size), the generated costumes have anatomical flaws (such as oversized heads and distorted hand proportions) owing to the single-task design. Textual Inversion introduces colour artefacts, especially the wrong shade of white in Ming costumes (figure 3, row 3), and unintentionally incorporates modern design elements. StyleCLIP suffers from severe detail degradation, producing blurred patterns and unnatural fabric transitions owing to unstable text

guidance. In contrast, our model successfully reproduces the iconic features of dynastic costumes through *V** embedding and LoRA optimisation.

Quantitative assessments

Text-image alignment analysis

The cross-model evaluation (table 2) shows different performance characteristics for the models. Although Textual Inversion achieves the highest text alignment, this comes at the cost of a severe drop in image alignment and an elevated KID/MMD score, suggesting that the erroneous presence of modern elements in the dynastic costumes resulted in semantic bias. Dreambooth demonstrates a balanced but suboptimal alignment. Our model achieves an optimal balance, achieving competitive text alignment along with excellent image fidelity while reducing the KID and MMD by 13% and 8%, respectively, compared to Dreambooth. This balance stems from the V*-guided attention mechanism, which dynamically weights dynasty-specific textual prompts during cross-attention operations, specifically enhancing the preservation of dynasty-specific structures and continuity of costume patterns.

Ablation experiments

The initialisation strategy analysis (table 3) highlights the critical trade-offs in feature anchoring. Fixed V* initialisation produced excellent text alignment but exhibited high KID and MMD owing to rigid semantic constraints that hindered adaptive feature fusion. The optimised V* initialisation matches the text alignment degree of the random initialisation, while there is an improvement in image alignment and a lower MMD, which effectively mitigates problems such as historical inaccuracy. The component contribution tests (table 4) show the necessity of a dual adaptation mechanism. The removal of LoRA reduces model alignment, with significant increases in the KID and MMD. The complete removal of V* and LoRA further exacerbates these problems, reflecting significant deviations from historical distribution. The results suggest that the synergy between the optimised V* (managing epoch-specific semantics) and LoRA (enabling parameter-efficient feature adaptation) is effective for reconstructing complex and specific details.

Table 2

CROSS-MODEL TEXT-IMAGE ALIGNMENT PERFORMANCE

Model	Text alignment ↑	Image alignment ↑	KID ↓	MMD ↓
Dreambooth	0.77	0.76	16.82	3.21
Textual inversion	0.80	0.69	18.95	4.05
Our model	0.79	0.78	14.64	2.96

Table 3

ABLATION STUDY ON V* INITIALIZATION STRATEGIES

Condition	Text alignment ↑	Image alignment ↑	KID ↓	MMD ↓
Random init	0.80	0.76	15.57	4.08
Fixed V*	0.83	0.75	16.44	5.06
Optimised V*	0.80	0.77	15.57	3.43

Table 4

CROSS-MODEL TEXT-IMAGE ALIGNMENT PERFORMANCE

Configuration	Text alignment ↑	Image alignment ↑	KID ↓	MMD ↓
Full model	0.79	0.78	14.64	2.96
w/o LoRA	0.75	0.76	15.57	3.43
w/o V* + LoRA	0.77	0.71	17.74	5.61

CONCLUSIONS

In conclusion, this study proposes an enhanced lightweight framework based on SD for generating historically accurate images of ancient costumes. Compared with the baseline model, the proposed method demonstrates excellent performance in restoring multi-dynasty costume features, especially in achieving higher historical fidelity in key details. The quantitative assessment further confirms the better match between the generated images and the reference historical sources. Despite the success of our method in generating ancient costumes, we recognise the main limitations that must be addressed. Primarily, the current dataset, while focused on core Han-style costumes, is limited in scope (encompassing only the Tang, Song, and Ming dynasties) and size (2,031 images). Limited by the scarcity of authoritative historical documents and the complexity of multi-dynastic costume form examinations, this scale remains small compared to large-scale benchmark datasets. Although we have adopted technical measures such as data augmentation, the risk of overfitting has not been eliminated in small sample fine-tuning. Furthermore, the labelling scheme treats each dynasty as a single entity and does not account for stylistic evolution within different periods of a dynasty, which may reduce the historical accuracy of the generated results. It is worth noting that data scarcity and detailed historical representation are common challenges in the field of textile artefact digitisation. In future research, we will prioritise the construction of an expanded dataset covering more dynasties and diverse ethnic costumes. At the same time, we will consider more detailed time labels (such as early/mid/late Tang Dynasty) to better capture the evolution within each dynasty. Additionally, we will develop a feature decoupling module for fabric features based on high-resolution 3D scanning and promote the application of digital reconstruction and innovative design of ancient costumes by relying on the multimodal semantic understanding advantages of models like DeepSeek Janus-Pro [29].

REFERENCES

[1] Ding, Q.-K., Liang, H.-E., *Digital restoration and reconstruction of heritage clothing: a review*, In: Herit Sci, 2024, 12, 1, 225, https://doi.org/10.1186/s40494-024-01349-4

[2] Liu, L., Zhang, H., Li, Q., Ma, J., Zhang, Z., *Collocated Clothing Synthesis with GANs Aided by Textual Information: A Multi-Modal Framework*, In: ACM Trans. Multimedia Comput. Commun. Appl., 2024, 20, 1, 1–25, https://doi.org/10.1145/3614097

[3] Wu, Q., et al., *ClothGAN: generation of fashionable Dunhuang clothes using generative adversarial networks*, In: Connection Science, 2021, 33, 2, 341–358, https://doi.org/10.1080/09540091.2020.1822780

[4] Wu, X., Li, L., *An application of generative AI for knitted textile design in fashion*, In: The Design Journal, 2024, 27, 2, 270–290, https://doi.org/10.1080/14606925.2024.2303236

[5] Goodfellow, I., et al., *Generative adversarial networks*, In: Commun. ACM, 2020, 63, 11, 139–144, https://doi.org/10.1145/3422622

[6] Radford, A. et al., *Learning Transferable Visual Models From Natural Language Supervision*, 2021, arXiv, https://doi.org/10.48550/ARXIV.2103.00020

[7] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D., *StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery,* 2021, arXiv: arXiv:2103.17249, https://doi.org/10.48550/arXiv.2103.17249

[8] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., *High-Resolution Image Synthesis with Latent Diffusion Models*, 2021, arXiv, https://doi.org/10.48550/ARXIV.2112.10752

[9] Avrahami, O., et al., *SpaText: Spatio-Textual Representation for Controllable Image Generation*, In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, 18370–18380, https://doi.org/10.1109/CVPR52729.2023.01762

[10] Saito, K., et al., *Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval*, 2023, arXiv, https://doi.org/10.48550/ARXIV.2302.03084

[11] Liu, K., Wu, H., Ji, Y., Zhu, C., *Archaeology and Restoration of Costumes in Tang Tomb Murals Based on Reverse Engineering and Human-Computer Interaction Technology*, In: Sustainability, 2022, 14, 10, 6232, https://doi.org/10.3390/su14106232

[12] Liu, K., Gao, Y., Zhang, J., Zhu, C., *Study on digital protection and innovative design of Qin opera costumes*, In: Herit Sci, 2022, 10, 1, 127, https://doi.org/10.1186/s40494-022-00762-x

[13] Shi, Z., Xiong, B., *Fine-Tuning Text-to-Image Generation Models Using Curriculum Learning for Yao Costume Image Generation*, In: 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China: IEEE, Mar. 2024, 1305–1311, https://doi.org/10.1109/AINIT61980.2024.10581585

[14] van den Oord, A., Vinyals, O., Kavukcuoglu, K., *Neural Discrete Representation Learning*, 2017, arXiv, https://doi.org/10.48550/ARXIV.1711.00937

[15] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A., *Generative Adversarial Networks: An Overview*, In: IEEE Signal Process. Mag., 2018, 35, 1, 53–65, https://doi.org/10.1109/MSP.2017.2765202

[16] Huang, Y., et al., *Diffusion Model-Based Image Editing: A Survey*, 2024, https://doi.org/10.48550/ARXIV.2402.17525

[17] Xiong, S., Pan, L., Ma, X., Hu, Q., Beckman, E., *Unsupervised deep hashing with multiple similarity preservation for cross-modal image-text retrieval*, In: Int. J. Mach. Learn. & Cyber., 2024, 15, 10, 4423–4434, https://doi.org/10.1007/s13042-024-02154-y

[18] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., *Hierarchical Text-Conditional Image Generation with CLIP Latents*, 2022, arXiv, https://doi.org/10.48550/ARXIV.2204.06125

[19] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.-Y., *Multi-Concept Customization of Text-to-Image Diffusion*, 2023, arXiv: arXiv:2212.04488, https://doi.org/10.48550/arXiv.2212.04488

[20] Ronneberger, O., Fischer, P., Brox, T., *U-Net: Convolutional Networks for Biomedical Image Segmentation*, In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Navab, N., Hornegger, J., Wells, W.M. and Frangi, A.F. Eds., Cham: Springer International Publishing, 2015, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28

[21] Vaswani, A. et al., Attention Is All You Need, 2017, arXiv, https://doi.org/10.48550/ARXIV.1706.03762

[22] Abdal, R., Qin, Y., Wonka, P., *Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?*, In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, 4431–4440, https://doi.org/10.1109/ICCV.2019.00453

[23] Hu, E.J., et al., *LoRA: Low-Rank Adaptation of Large Language Models*, 2021, arXiv, https://doi.org/10.48550/ARXIV.2106.09685

[24] Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y., *CLIPScore: A Reference-free Evaluation Metric for Image Captioning*, 2021, https://doi.org/10.48550/ARXIV.2104.08718

[25] Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., *Demystifying MMD GANs,* 2018, arXiv, https://doi.org/10.48550/ARXIV.1801.01401

[26] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., *A kernel two-sample test*, In: J. Mach. Learn. Res., 2012, 13, 723–773

[27] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K., *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation*, In: 2023 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, 22500–22510, https://doi.org/10.1109/CVPR52729.2023.02155

[28] Gal, R., et al., *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion*, 2022, arXiv, https://doi.org/10.48550/ARXIV.2208.01618

[29] Chen, X., et al., *Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling*, 2025, 2025, arXiv: arXiv:2501.17811, https://doi.org/10.48550/arXiv.2501.17

Authors:

JIAN HUA[1], ZHI LI[2], RUIHONG CHEN[2], YU CHEN[2]

[1]School of Textiles and Fashion, Shanghai University of Engineering Science, China

[2]Shanghai University of Engineering Science, 333 Longteng Road, Shanghai 201620, China

Corresponding author:

YU CHEN
e-mail: ychen0918@sues.edu.cn