International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 2321-2332

https://doi.org/10.70517/ijhsa463194

A Study of Controllability Paths for Generative AI Design Based on Stable Diffusion ControlNet

Chen Zou[1] and Yanyan Chen[2,*]
[1] XianDa College of Economics and Humanities, Shanghai International Studies University, Shanghai, 202162, China
[2] Guilin Normal College, Guilin, Guangxi, 541000, China
Corresponding authors: (e-mail: chen_chen1323@163.com).

Abstract With the continuous iteration of algorithms in the field of deep learning, generative AI design is ushering in a revolutionary change. In this project, we study the controllability path of generative AI, take Stable Diffusion as the base model of image generation, use the improved LoRA method and Controlnet to fine-tune its control, and realize the image generation method based on Stable Diffusion ControlNet. The experimental results show that the image quality of the image generation model designed in this paper obtains better performance in IS, FID and other evaluation indexes with reasonable scale parameter and sampling step size. 8.83% and 76.07% of IS and FID values are improved compared with the Stable Diffusion1 model when scale parameter is set to 8, and 1000 sampling steps are used to achieve better image quality than the Stable Diffusion1 model. The minimum FID value of 1.18 is obtained when the sampling step is 1000, which verifies the effectiveness of the Stable Diffusion ControlNet network designed in this paper. The silk scarf pattern generated by the model scores 5.36 and 5.47 in artistic aesthetics and normality, and the generated pattern is clear and of excellent quality. In addition, the model meets the requirements of practical applications in terms of computational efficiency and hardware cost. The results show that the proposed Stable Diffusion ControlNet model can be used as a generative AI method with good fine-tuning.

Index Terms stable diffusion, Controlnet, DyLoRA, image generation, generative AI

I. Introduction

In the context of today's digital era, Artificial Intelligence Generative (AIGC) technology has penetrated into a wide range of fields. Initially, this technology was mainly applied to text generation, but today, its application scope has been expanded to image creation, audio processing, 3D modeling and other dimensions, which has had a profound impact on the art and design industry that uses digital means of creation [1], [2]. Relying on advanced data processing modes such as neural network modeling, generative AI technology is able to respond quickly to commands and generate brand new image content, which greatly improves creative efficiency and significantly reduces the cost of design time [3]-[5]. In the current social context of rapid information exchange and development, how to efficiently utilize AI design tools in order to enhance the creative ability and save time costs has become an important issue that designers must face [6], [7].

In the continuous evolution of generative AI technology, a large number of drawing tools have emerged in the field of AI design [8]. These advanced AI drawing tools construct a huge image database by parsing the input data and are able to generate brand new images matching the user's instructions, and their core competitiveness lies in the computing mechanism and data model behind them [9]-[11]. Among them, Stable Diffusion is an open-source AI painting tool, which works on the principle of destroying the training data by successively adding Gaussian noise, and then reversing this noise process as a way to recover the data [12]. After training, the model is able to synthesize new data from random inputs, enabling algorithmic innovation [13]. Although Stable Diffusion can describe an image in natural language and generate an image that matches that description, the generated image is highly uncertain and has some limitations [14], [15].

The Stable Diffusion mapping model has a very rich resource of models and extension plug-ins in addition to these main features such as text-born maps and graph-born maps [16]. The emergence of ControlNet plug-in introduces a new control method for the image generation function, which can control the screen in a more stable way, which is a good solution to the problem that the text-generated image cannot control the details of the generated image [17], [18]. Painting using Stable Diffusion, along with ControlNet and other plug-ins, can easily generate the expected picture, and even if the creator has no art skills, he or she can quickly create works with complex gestures and multiple painting styles [19]-[21].

In this paper, based on Stable Diffusion and its controllability, we improve the dynamic low-rank adaptive algorithm, propose a new parameter decomposition method to improve the rank of the parameter matrix, so as to realize the fine-tuning of the stable diffusion model, combine the semantic information of the Stable Diffusion model with the bearer DyLoRA module, and control the image generation by using the Controlnet to construct the image generation method based on Stable Diffusion ControlNet. Choose to conduct experiments on Fashion-MNIST dataset to explore the image generation quality of the model under different scale parameters and sampling steps using IS and FID values as evaluation indexes. Then take the flower silk scarf pattern generation as an example to carry out subjective evaluation experiments, and collect the questionnaire data of the images generated by this paper's method and the comparison method in terms of artistic aesthetics and silk scarf specification, and analyze the subjective evaluation results of the images generated by this paper's method. We also explore the computational efficiency and hardware cost of this paper's method from the dimensions of training time, inference time and video memory occupation.

II. Stable Diffusion and its controllability

Generative AI design involves a variety of software and models, such as Stable Diffusion, Disco Diffusion and MidjourneyV5. They have different characteristics, as evidenced by different algorithms and deployment methods, open source situations, etc. Stable Diffusion is lightweight and has high output quality, and its special algorithms make the arithmetic requirements of Stable Diffusion relatively low. With the support of plug-ins, the images generated by Stable Diffusion are more controllable and can meet the needs of designers, and have the potential to be widely used in the design industry.

Stable Diffusion's image generation has great randomness, which can be gradually stabilized by controlling the parameters. Through the designer's human correction, it can save a lot of time delayed because of random generation, so as to achieve the global optimization of efficiency. The main parameters that control SD are:

II. A. Render steps

The number of rendering steps (steps) is an important parameter that affects the imaging quality in SD.A generalized description of the SD imaging process is the process of cyclic noise diffusion, and the default number of rendering steps in the initial settings of SD is 20 steps. Under a fixed seed, the more rendering steps the figure has, the richer the imaging details become, and as the number of steps gets larger and larger, the imaging details tend to be saturated, and the image is gradually stabilized, with the largest change from 0-20 steps.

II. B. Sampling method

Sampling method, i.e., the de-noised sampling pattern of a diffusion algorithm. In generative AI design, sampling approach refers to the technique of selecting and extracting information from an input image or dataset for the purpose of generating or controlling the image with an AI algorithm. It plays a crucial role in determining how an AI system perceives and interprets visual content. Sampling methods involve dividing the input image into pixels or smaller units and analyzing them individually or in groups.AI algorithms analyze the colors, textures, shapes, and other visual features present in these blocks of pixels in order to learn patterns and relationships in the dataset.

II. C. Picture specifications

With digital images in pixels, a larger image specification means a larger block of pixels. However, in generative AI design, with the same number of seeds, the image has different specifications and the content of the image is also different. As the size of the diagram changes, the content of the image also changes, the larger the specification of the image, the more elements it contains.

II. D. CFG scale

CFG scale is a parameter that affects the weights of words in generative AI design.CFG scale refers to the degree of agreement between the image and the cue - the higher it is, the closer the generated graph is to the words described by the user. However, in practice, the value should be set within a reasonable range, not the higher the better, if the value is set too high, even overfitting may occur.

II. E. Diagram Generation

Graph born diagram is another mode of SD, which is to convert one kind of image to another by image conversion technique, suitable for teams with art development ability. Redraw Amplitude is an important parameter of SD graph-generated graph, which determines how much the algorithm affects the content of the image. When set to 0, the output image is unchanged. When set to 1, a completely different image is generated, and different modeling styles will also affect the results.

III. Stable Diffusion ControlNet based image generation

Based on the controllability of Stable Diffusion, this paper adds LoRA fine-tuning method to Stable Diffusion and introduces ControlNet to control it, to design a controllability path for generative AI design based on Stable Diffusion ControlNet, and to obtain an improved Stable Diffusion's image generation model.

III. A. General Control Framework for Stable Diffusion

Stable Diffusion itself can only receive text encoding, time encoding (scheduler), and latent space noise. The time coding (scheduler) has EulerA, DDIM, DPM, etc., which affects the inference speed, but all of them have a small impact on the generated results, so they are not listed as a dimension alone. The structure of Stable Diffusion network after adding LoRA, ControlNet fine-tuning is shown in Figure 1. In this framework, the generation process of Stable Diffusion can be controlled by five dimensions, the big model, cue words, latent space noise, ControlNet and LoRA.

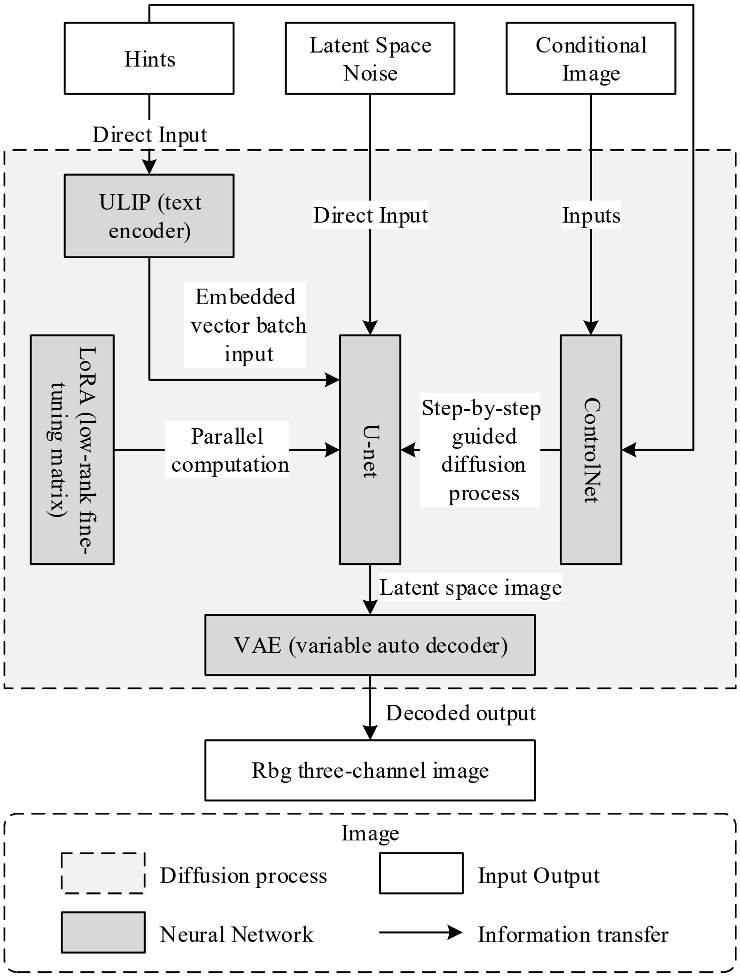

Figure 1: Stable Diffusion under multiple control methods

III. B. Stable Diffusion Model

Stable Diffusion is a model based on a latent diffusion model structure, pre-trained by StabilityAI.Stable Diffusion is mainly composed of three core components: CLIP Text Encoder, U-Net, and VAE.

III. B. 1) CLIP model

Stable Diffusion uses the CLIP text encoder to transform cue words into semantic vectors.The CLIP model is a model developed by OpenAI that aligns images and text into a shared semantic space, making it possible to derive the relevant text from an image, as well as to find the corresponding image from a textual description.

CLIP is trained by contrast learning. It does this by using a large number of image-text pairs as positive examples and random image-text pairs as negative examples, and then training the model so that the text and images of the positive examples are closer together in the model's mapping space, and the negative examples are farther away.The CLIP model consists of two parts: an image encoder and a text encoder. The image encoder is a pre-trained convolutional neural network and the text encoder is a pre-trained Transformer model. To encode the text, CLIP first uses a word embedding layer to convert the text into an embedding vector, and then feeds this vector into the Transformer model, which converts this embedding vector into a semantic vector by means of a self-attention mechanism and positional encoding. To invert the cue word from the image, the CLIP model encodes the image as a vector and then finds the text vector closest to this vector in the shared semantic space. This is achieved by minimizing the distance between the image vector and the corresponding text vector during training.

The text encoder part of the CLIP model is mainly used in Stable Diffusion. All the cue words in the input are converted into converted into text encoding of size 77*768 by CLIP.

III. B. 2) U-Net

The U-Net network structure in Stable Diffusion has a total of 25 layers, 12 layers for encoding, 1 layer for intermediate and 12 layers for decoding.The U-Net model is divided into two parts: encoder and decoder. The main responsibility of the encoder module is to perform feature extraction and learning. The decoder module's responsibility is to restore the feature mapping to the original resolution. A key step in this process is the fusion of shallow positional information with deep semantic information through skip-connection.

In Stable Diffusion, the key core of the diffusion model is the U-Net model, whose main task is to denoise the latent space noise step by step. The latent space noise data is looped from U-Net several times for iterative denoising: the U-Net has been trained to predict the noise based on the current latent space image x_t and each time the predicted noise is guided by the input cue words i.e., the text encoding z_t and the time encoding t. The predicted noise is removed on x_t. The iterated x_{t-1} is used as a new input to the U-Net. After multiple passes the latent space image is output to the VAE for decoding and reconstruction into a pixel level image.

Each coding and decoding layer of U-Net consists of a residual network module, a cross-attention module, and a down- or up-sampling module. The cross-attention module, which is used in Stable Diffusion U-Net, is the Spatial Transformer, which mainly implements the direction in which the text encoding guides the noise prediction. This is the key means of controlling the content generated by the Stable Diffusion model.The Spatial Transformer module is able to integrate semantic information (i.e., direct a region on the image to generate the corresponding kind of textual content) at the corresponding location on the image without changing the size of the inputs and outputs.The Spatial Transformer can both be Spatial Transformer can be regarded as a module that connects text to images.

III. B. 3) VAE

The main role of VAE (Variable Auto-Encoder) in Stable Diffusion is a scaler of latent space images and real images. A Variational Autoencoder is a generative model that learns the process of generating data by means of probabilistic statistics and then generates new data based on this process.VAE has two important components: an encoder and a decoder. The encoder maps data from a high-dimensional input space to a low-dimensional hidden space, which is usually assumed to obey some probability distribution (usually Gaussian). The decoder, in turn, maps the data from the hidden space back to the high-dimensional input space.

During the Stable Diffusion training process, the real image is converted into a latent space image by the VAE encoder.The VAE uses a relative downsampling factor of 8 and maps an image of shape XxWx3 to a latent space image of shape H/8*W/8*4. And in the generation stage, VAE's decoder converts the U-Net noise-removed latent space image into an RGB three-channel image.

III. C. Model fine-tuning methods

The stable diffusion model is a larger network model, and training this model usually requires a larger amount of data and training costs. It is very difficult for training a stable diffusion model that is specialized in generating a certain type of image, so it is fine-tuned with LoRA and ControlNet.

III. C. 1) LoRA fine-tuning methods

LoRA freezes the weights of the pre-trained model and injects trainable layers (called rank decomposition matrices) in each Transformer block. The number of training parameters required is greatly reduced, thus lowering GPU memory requirements.The structure of LoRA is shown in Fig. 2.

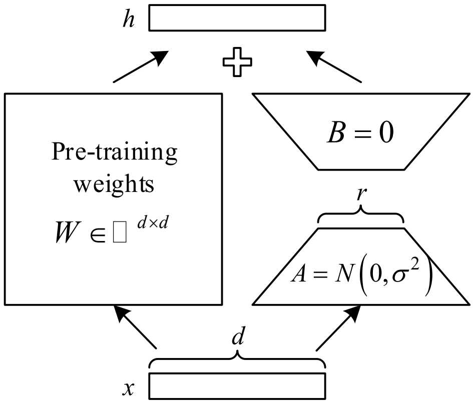

Figure 2: LoRA structure

There are many dense parameter matrices during the fine-tuning training process of the large model, but the change matrix of the parameters during the fine-tuning process is not full of rank, which can be changed to train a side branch by freezing the weights of the pre-trained model, such as the input x in the figure, which has the dimension d, and in the LoRA method, the data is firstly downgraded from the dimension of d to the dimension of r by using the linear layer A, which r is the LoRA s rank, which is one of the most important hyperparameters in LoRA. Generally r will be much smaller than d, followed by a second linear layer B, which changes the data from r-dimensional back to d-dimensional. Finally, the results of the left and right parts are added together to obtain a new intermediate vector h. In this process the number of parameters to be learned changes from $d\times d$ to $d\times r+r\times d$, which is greatly reduced in the case of $r \ll d$. Expressed in Eq:

$$h=W_0x+\Delta Wx=W_0x+\frac{\alpha}{r}W_{up}W_{dw}x \tag{1}$$

where: W_0 freezes the parameter matrix, W_{dw} and W_{up} are the parameter matrices of Linear layer A and B respectively, x is the input, r is the rank and α is the scaling factor.

However, in different tasks, the fixed rank often cannot meet the needs, and finding a suitable hyperparameter rank will waste a lot of resources. In this paper, an improved dynamic low-rank adaptation technique, DyLoRA, is used instead of ordinary LoRA.The basic idea of DyLoRA is the same as that of ordinary LoRA, but the implementation details have been greatly improved. Meanwhile, for the important hyperparameter rank, DyLoRA uses a range $\left[r_{\min},r_{\min+1},\cdots,r_{\max}\right]$ to replace the original rank r. The $r_{\min}$ and $r_{\max}$ are two new hyperparameters. During the training process a random sample b is taken from the well-fixed probability distribution $b\sim P_B$, $b\in\left\{r_{\min},r_{\min+1},\cdots,r_{\max}\right\}$ and accordingly truncated W_{dw}, W_{up} to obtain $W_{up\downarrow b}$ and $W_{dw\downarrow b}$, and the shapes of the decomposed two matrices become $d\times b$ and $b\times d$. The calculations are as in Eqs. (2) and (3):

$$W_{up\downarrow b}=W_{up}\left[:,1:b\right] \tag{2}$$

$$W_{dw\downarrow b}=W_{dw}\left[1:b,:\right] \tag{3}$$

Thus equation (1) is transformed in the forward process:

$$h=W_0x+\Delta Wx=W_0x+\frac{\alpha}{b}W_{up\downarrow b}W_{dw\downarrow b}x \tag{4}$$

While in the inverse process, the parameters are updated only for the b th column/row in order to prevent the loss of previously learned information, the other parameters are fixed.

The traditional DyLoRA parameter decomposition method has no positive increasing effect on the rank of the parameter matrix ΔW, but the size of the rank directly determines the amount of information the parameter matrix can carry. In this paper, DyLoRA is improved on the parameter matrix decomposition way, and for the parameter matrix ΔW it is divided into two sets of dot products of $W_{up\downarrow b}$, $W_{dw\downarrow b}$.

According to the new parameter matrix decomposition, the parameter matrix in the forward process can be expressed as:

$$h = W_0 x + \frac{\alpha}{b}\left(W_{up\downarrow b1} W_{dw\downarrow b1} \square\ W_{up\downarrow b2} W_{dw\downarrow b2}\right)x \tag{5}$$

where $b1 = b2 = b/2$, the rank of the left matrix is $b1$ and the rank of the right matrix is $b2$.

For dot product operations on matrices, there is $r(A\square\ B) \le r(A)r(B)$, and let A, B matrices be $m \times n$ matrices, $r(A) = r$, $r(B) = s$. Then A, B matrices can be written as $A = \sum_{i=1}^{r} \alpha_i \beta_i^T$, $B = \sum_{i=1}^{s} \gamma_i \beta_i^T$, and so we can get Eq. (6):

$$\begin{aligned} A\square\ B &= \sum_{i=1}^{r}\sum_{i=1}^{s}\left(\alpha_i \beta_i^T\right)\square\ \left(\gamma_i \eta_i^T\right) \\ &= \sum_{i=1}^{r}\sum_{i=1}^{s}\left(\left(\alpha_i \square\ \gamma_i\right)\left(\beta_i \square\ \eta_i\right)^T\right) \end{aligned} \tag{6}$$

where: $\alpha_1, \cdots, \alpha_i$ and $\beta_1, \cdots, \beta_i$ are m-dimensional linearly uncorrelated column vector sets, and $\gamma_1, \cdots, \gamma_i$ and $\eta_1, \cdots, \eta_i$ are n-dimensional linearly uncorrelated column vector sets. According to Eq. (6), Eq. (7) can be obtained:

$$\begin{aligned} r\left(A\square\ B\right) &= \sum_{i=1}^{r}\sum_{i=1}^{s} r\left(\left(\alpha_i \square\ \gamma_i\right)\left(\beta_i \square\ \eta_i\right)^T\right) \\ &\le r(A)r(B) = rs \end{aligned} \tag{7}$$

If the parameter decomposition is done in the way of the DyLoRA forward process, the two matrices decomposed are simply done as a cross-multiplication, and the rank of the multiplication of the two can only be less than or equal to the minimum value of the two matrices decomposed.

$$r\left(A \times B\right) \le \min\left\{r(A), r(B)\right\} = \min\left\{r, s\right\} \tag{8}$$

Ideally, the rank of the ΔW parameter matrix obtained after dot product is $b1 \times b2 = b^2/4$. If the parameter decomposition is performed in the manner of the DyLoRA forward process, the rank of the ΔW parameter matrix is b. Obviously when the value of b is larger $\left(b > 4\right)$, if the new parameter decomposition in the improved DyLoRA forward process is followed the ΔW parameter matrix can be made to have a larger rank with the same total number of parameters, thus obtaining more information.

As the parameter matrix is decomposed into 4 matrices in the inverse process, the number of parameters to be updated in each parameter matrix in the parameter update process is doubled compared to DyloRA, which is $4 \times d$, but the increase in the number of parameters is still insignificant compared to the updating of the whole large model $d \times d$, as its dimension d is much larger than 4.

III. C. 2) ControlNet

The diffusion model has achieved a high level of success in the field of "text-to-graph". For the same text cue, different random seeds can be used to generate different images, which is a reflection of its high versatility. But when you want to accurately control the layout of an image or the shape of an object, it is not enough to use only the textual cue, Controlnet can effectively solve this problem by adding additional conditions to control the generation of images.

The structure of Controlnet to control a single neural network block is shown in Figure 3. When Controlnet is not added, the original neural network F of the diffusion model inputs x to get y, and the parameters are represented by Θ:

$$y = F\left(x; \Theta\right) \tag{9}$$

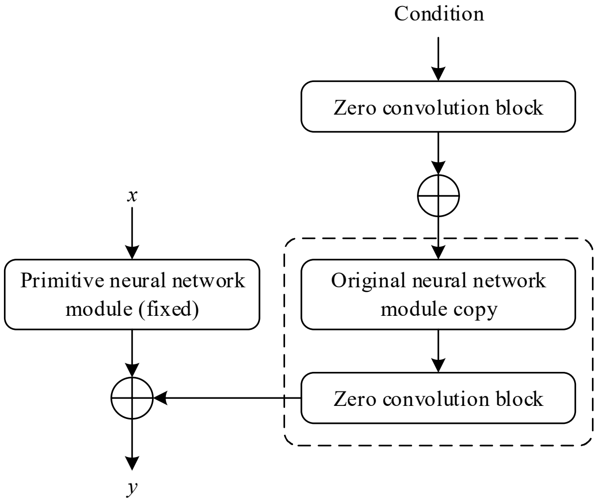
Condition
Zero convolution block
x
Primitive neural network module (fixed)
Original neural network module copy
Zero convolution block
y

Figure 3: Schematic diagram of ControlNet controlling a single neural network block

In ControlNet, the Encoder of the Unet of Stable Diffusion is locked, and then a copy is made to operate on the original neural network module copy to impose control conditions. The results after applying the control conditions were added to the results of the original model to obtain the final output. Finally, the output of the original network is modified:

$$y_c = F(x;\Theta) + Z\left(F\left(x + Z(c;\Theta_{z1});\Theta_c\right);\Theta_{z2}\right) \tag{10}$$

where: Z denotes the zero convolution, which is a 1×1 convolution with initialized weights and bias 0, Θ_{z1}, Θ_{z2} are the parameters of the two-layer zero convolution, and c is the condition of the input. After passing the control condition through the zero convolution, it is added to the original input, and the sum is entered into the replicated neural network block of ControlNet, and the output of the network is added to the output of the original network after doing another zero convolution. The initial state parameters of the untrained ControlNet are given in Eqs. (11)-(13) below:

$$Z(c;\Theta_{z1}) = 0 \tag{11}$$

$$F\left(x + Z(c;\Theta_{z1});\Theta_c\right) = F(x;\Theta) \tag{12}$$

$$Z\left(F\left(x + Z(c;\Theta_{z1});\Theta_c\right);\Theta_{z2}\right) = 0 \tag{13}$$

When Controlnet is untrained, the output is 0, and the initial state added to the original network is also 0. This has no effect on the original network, ensuring that the performance of the original network is preserved intact.

The above implements Controlnet's control of a single neural network block, while the process of controlling the entire stabilized diffusion network is to replicate the entire encoder in it for training, with the decoder portion jump-connected. Its loss function is shown in equation (14):

$$L = \mathrm{E}_{z_0,t,c_t,c_f}\left[\left\|\varepsilon - \varepsilon_\theta(z_t,t,c_t,c_f)\right\|^2\right] \tag{14}$$

where ε_θ represents the network, t denotes the time step, c_t denotes the textual control, and c_f denotes the control condition.

IV. Experimental process and result analysis

In order to verify the effectiveness of the Stable Diffusion ControlNet-based generative AI designed in the previous chapter for designing controllability paths, the basic Stable Diffusion model and Stable Diffusion ControlNet model were used for training, respectively, and the control group chose ADM, CGAN, ASG- GAN and other image generation models, and subjective evaluation experiments were conducted to analyze the model-generated images.

IV. A. Experimental data set

The experiments are trained using Fashion-MNIST dataset.Fashion-MNIST consists of a training set of 60,000 examples and a test set of 10,000 examples.Each example is a 28×28 grayscale image divided into a total of 10 categories, which are different from the MNISTQ handwritten dataset.Fashion-MNIST dataset contains images in 10 categories, namely: t-shit, touser, pulover, dress, coat, sandal, shit, sneaker, bag, ankleboot.Fashion-MNIST is a rich, well-structured and easy-to-use dataset that is well suited for image classification, image recognition and other computer vision tasks for research and experimentation.

IV. B. Analysis of experimental results

(1) Basic control group

The Stable Diffusion model was trained using the Fashion-MNIST dataset, the input image size was 28×28, the number of channels was 1, the noise_step was set to 1000, the batchsize was set to 128, the scale was firstly set to 10, the training step epoch was set to 20, and the optimizer used the Adam algorithm, and the learning rate is set to 0.01. The loss function is plotted using TensorBoard as shown in Figure 4. It can be seen that the Stable Diffusion model generation is still slow, and the loss function reaches saturation after the training step size reaches 20k, and cannot continue to decrease with training.

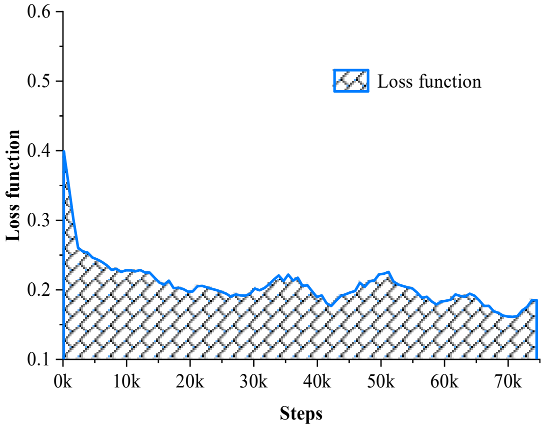

Figure 4: Loss function of unconditional control Stable Diffusion

(2) Control experiment results

In order to study the effect of different strengths of classification noise on the generator, different sizes of scale parameters were set for training, with scale set to 0, 2, 4, and 8, respectively, and the rest of the conditions remained unchanged. ADM, CGAN, and ASG-GAN were chosen as control groups. The IS and FID parameters were selected as the evaluation criteria for judging the quality of the output images, and the comparison of the quality of the images generated by different methods is shown in Table 1.The larger the scale parameter of the Stable Diffusion ControlNet model is, the better the IS and FID indices of the images are, and it can be concluded that, if the scale parameter is set reasonably, the Stable Diffusion ControlNet method used in this paper Stable Diffusion ControlNet method can achieve better image generation quality. When the scale parameter is 8, the IS and FID indices of the image are 91.30 and 1.23, which are significantly better than the other methods, and the IS index is improved by 18.83% and the FID index is reduced by 76.07% compared with Stable Diffusion.

(3) Further experiments In order to study the effect of different sampling step lengths on the objective evaluation quality of the output image of the Stable Diffusion ControlNet model, the scale parameter is set to be unchanged at 8, the training set adopts FashionMNIST, and the evaluation parameters adopt IS and FID, and the Fashion-MNIST training is performed with different sampling step lengths. The experimental results are shown in Fig. 5. After the sampling step size of noise_steps exceeds 150 steps, the IS value of the generated effect tends to stabilize at around 95. When the sampling step size is 1000, the FID obtains the minimum value of 1.18. The experimental results show that when using the Stable Diffusion ControlNet model in this paper for training, a better output image quality can be obtained when the sampling step size is set to 1000 steps.

Table 1: The quality contrast of the image generated by different methods

Models	scale	IS ↑	FID ↓
Stable Diffusion ControlNet	0	40.79	2.48
	2	33.21	3.79
	4	67.45	1.88
	8	91.30	1.23
Stable Diffusion	-	76.83	5.14
ADM	-	89.68	5.67
CGAN	-	84.81	4.14
ASG-GAN	-	87.37	4.11

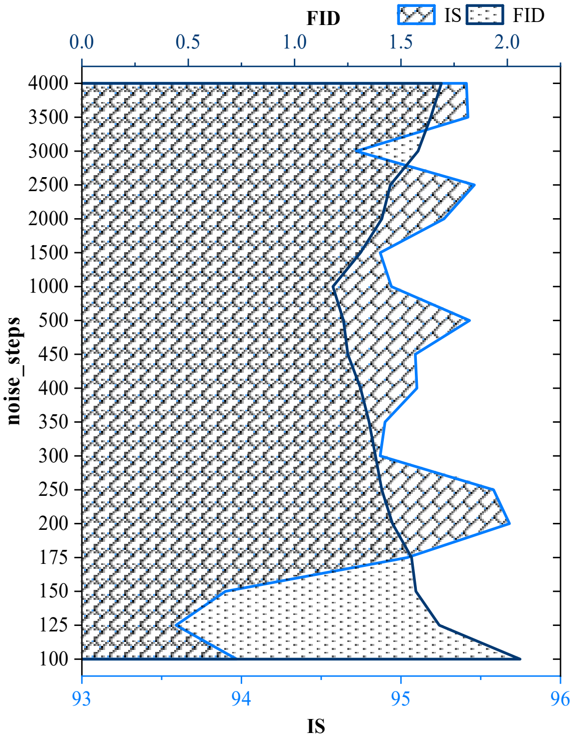

Figure 5: Training results of Fashion-MNIST under different noise_steps

IV. C. Subjective assessment experiments

In order to comprehensively assess the performance of the image generation methods in this paper from a subjective point of view, taking the generation of floral silk scarf patterns as an example, and using the five methods above to generate relevant images, a questionnaire containing five questions corresponding to the five image generation methods is designed. The questionnaire is a subjective assessment of the finely drawn floral silk scarf pattern from the dimensions of artistic aesthetics and silk scarf standardization. Artistic aesthetics refers to the degree to which the generated pattern is visually aesthetically pleasing, giving a sense of pleasure and attractiveness. Silk scarf normality refers to the degree to which the generated pattern conforms to the real silk scarf pattern in terms of composition and border design, which is used to measure whether the generated pattern follows the design norms and characteristics of the real silk scarf pattern, and a high silk scarf normality implies that the generated pattern is visually closer to the real silk scarf design, and conforms to the actual standard of use. The questionnaire was rated on a 5-point scale, with 1 indicating the lowest rating and 5 indicating the highest. A total of 128 valid questionnaires were recovered from this questionnaire survey, and the statistical results of the questionnaire are shown in Figure 6.

In this paper, the Stable Diffusion ControlNet model has a balanced performance in artistic aesthetics and silk scarf specification, and the proportion of the number of people choosing 4 and 5 points in both dimensions is more than 82%, which shows a higher degree of recognition.The CGAN and ASG-GAN models, although the proportion of the ratings of 4 and 5 points is also higher, the proportion is higher than that of the Stable Diffusion ControlNet model is reduced, the proportion of 4 and 5 points in artistic aesthetics and normality is 60.94% and 57.81% for CGAN model, and 63.28% and 59.38% for ASG-GAN model, the overall recognition is not as good as that of Stable Diffusion ControlNet model.The ADM method scores show an intermediate centralized tendency, the The number of choices with scores of 2 and 3 is relatively high, showing a weak polarization but low overall recognition.The Stable Diffusion model performs the worst, with an over-representation of scores of 1 and 2 for both dimensions, which is clearly behind the other methods.

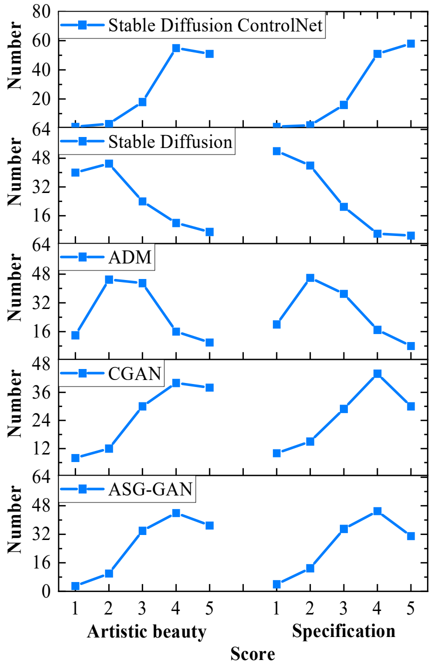

Figure 6: Statistical results of the questionnaire

Figure 7 shows the quantitative scores of the different methods for generating silk scarf patterns, which were calculated by averaging all the questionnaire results.The Stable Diffusion model achieved the lowest scores on artistic aesthetics and silk scarf norms, with scores of 2.85 and 2.52, respectively, indicating that the patterns it generated performed poorly in terms of artistic aesthetics and silk scarf norms.The Stable Diffusion ControlNet model achieves the highest scores in both artistic aesthetics and silk scarf standardization, with 5.36 and 5.47 points, respectively, indicating that it has a significant advantage in generating fine-drawn floral silk scarf patterns with high artistic aesthetics and in compliance with the specifications, and it can satisfy the requirements of artistic aesthetics and silk scarf standardization at the same time. In summary, it can be seen that the Stable Diffusion ControlNet model generates the best image quality among these methods.

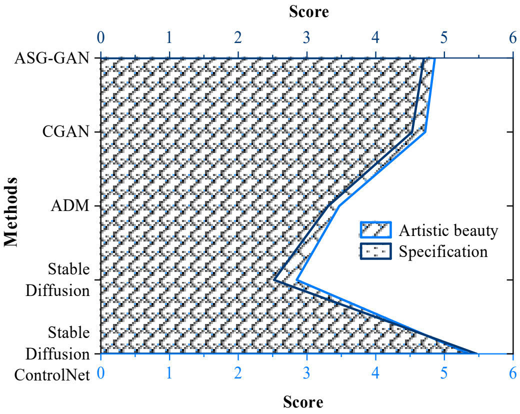

Figure 7: Quantitative scores of silk towel patterns generated by different methods

IV. D. Computational efficiency and hardware costs

In order to evaluate the computational efficiency and hardware cost of the methods in this paper, the training time, inference time, and graphic memory occupation of different methods under the same hardware environment are recorded. The computational efficiency and hardware cost of different methods are shown in Table 2.The Stable Diffusion ControlNet model outperforms the ADM model in three aspects: training time, inference time, and memory occupation, which indicates that the Stable Diffusion ControlNet model outperforms the ADM model in terms of computational efficiency and hardware cost. The training time, inference time, and video memory occupation of the Stable Diffusion ControlNet model slightly increase based on the benchmark model Stable Diffusion, with the training time increasing by 4h, or 13.33%, and the inference time increasing by 0.002s, or 6.67%, which is a smaller increase. In summary, the Stable Diffusion ControlNet model used in this paper meets the requirements of practical applications in terms of computational efficiency and hardware cost.

Table 2: Computational efficiency and hardware cost of different methods

Models	Training time/h	Inference time/s	Memory occupancy/GB
Stable Diffusion ControlNet	34	0.032	2.8
Stable Diffusion	30	0.030	2.7
ADM	36	0.036	3.2
CGAN	17	0.047	4.4
ASG-GAN	25	0.031	3.4

V. Conclusion

For generative AI design, this paper proposes a controllability path based on Stable Diffusion Controlnet by using DyLoRA and Controlnet to control the generated content. Experiments are conducted using Fashion-MNIST dataset, the IS and FID indices of images generated by the Stable Diffusion Controlnet model increase with the increase of scale parameter, and the IS and FID indices are 91.30 and 1.23 for scale parameter 8, which are better than Stable Diffusion1 model 8.83% and 76.07%. In addition, the FID of the model achieves the minimum when the sampling step is 1000, which is 1.18. The generated image quality of the model can be improved by reasonable parameter design. In the subjective evaluation experiments, the percentage of this paper's Stable Diffusion ControlNet model scoring 4 and 5 on the two evaluation metrics is more than 80%, with scores of 5.36 and 5.47, which is higher than the comparison methods. In terms of computational efficiency and hardware cost, the Stable Diffusion ControlNet model is slightly inferior to the Stable Diffusion model, but still meets the practical needs. Experiments show that the Stable Diffusion ControlNet method exhibits obvious advantages in image generation quality.

In this paper, facing the current image generation model based on the Stable Diffusion model, the solution of DyLoRA and Controlnet is proposed, which controls the image generation of the model and improves the visual perception quality of the output image, but the model design in this paper still has certain deficiencies, and there is a gap between the computational efficiency and hardware cost and the current mainstream commercial model, which needs to be further research.

Funding

This work was supported by 2024 University-Level Scientific Research Projects of XianDa College of Economics and Humanities (A3107.24.1801.2460).

References

[1] El Ardeliya, V., Taylor, J., & Wolfson, J. (2024). Exploration of artificial intelligence in creative fields: Generative art, music, and design. International Journal of Cyber and IT Service Management, 4(1), 40-46.

[2] Archana Balkrishna, Y. (2024). An analysis on the use of image design with generative AI technologies. International Journal of Trend in Scientific Research and Development, 8(1), 596-599.

[3] Hughes, R. T., Zhu, L., & Bednarz, T. (2021). Generative adversarial networks–enabled human–artificial intelligence collaborative applications for creative and design industries: A systematic review of current approaches and trends. Frontiers in artificial intelligence, 4, 604234.

[4] Messer, U. (2024). Co-creating art with generative artificial intelligence: Implications for artworks and artists. Computers in human behavior: artificial humans, 2(1), 100056.

[5] Hutson, J., & Harper-Nichols, M. (2023). Generative AI and algorithmic art: disrupting the framing of meaning and rethinking the subject-object dilemma. Global Journal of Computer Science and Technology: D, 23(1).

[6] Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. PNAS nexus, 3(3), pgae052.

[7] Guridi, J. A., Cheyre, C., Goula, M., Santo, D., Humphreys, L., Souras, A., & Shankar, A. (2025). Image generative ai to design public spaces: a reflection of how ai could improve co-design of public parks. Digital Government: Research and Practice, 6(1), 1-14.

[8] Chacón, J. C., Nimi, H. M., Kloss, B., & Kenta, O. (2021). Towards the development of AI based generative design tools and applications. In Design, Learning, and Innovation: 5th EAI International Conference, DLI 2020, Virtual Event, December 10-11, 2020, Proceedings 5 (pp. 63-73). Springer International Publishing.

[9] Saadi, J., & Yang, M. (2023). Observations on the implications of generative design tools on design process and designer behaviour. Proceedings of the Design Society, 3, 2805-2814.

[10] Adhikari, M. S., Verma, Y. K., Sindhwani, M., & Sachdeva, S. (2025). Generative AI Tools for Product Design and Engineering. Generative Artificial Intelligence in Finance: Large Language Models, Interfaces, and Industry Use Cases to Transform Accounting and Finance Processes, 299-325.

[11] Casteleiro-Pitrez, J. (2024). Generative artificial intelligence image tools among future designers: A usability, user experience, and emotional analysis. Digital, 4(2), 316-332.

[12] Khlewee, I. K. (2025). Image Generation Using Generative AI: Comparison Between OpenAI Art and Stable Diffusion. Babylonian Journal of Artificial Intelligence, 2025, 15-22.

[13] Dehouche, N., & Dehouche, K. (2023). What's in a text-to-image prompt? The potential of stable diffusion in visual arts education. Heliyon, 9(6).

[14] Zhang, X. (2024). AI-Assisted Restoration of Yangshao Painted Pottery Using LoRA and Stable Diffusion. Heritage (2571-9408), 7(11).

[15] Kabir, A. I., Mahomud, L., Al Fahad, A., & Ahmed, R. (2024). Empowering Local Image Generation: Harnessing Stable Diffusion for Machine Learning and AI. Informatica Economica, 28(1), 25-38.

[16] Rahmatulloh, A. (2024). Custom Concept Text-to-Image Using Stable Diffusion Model in Generative Artificial Intelligence. JICO: International Journal of Informatics and Computing, 1(1), 1-11.

[17] Zhao, S., Chen, D., Chen, Y. C., Bao, J., Hao, S., Yuan, L., & Wong, K. Y. K. (2023). Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems, 36, 11127-11150.

[18] Hartley, Z. K., Lind, R. J., Pound, M. P., & French, A. P. (2024). Domain targeted synthetic plant style transfer using stable diffusion LoRA and ControlNet. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5375-5383).

[19] Deng, B., & Lu, Y. (2025). Weed image augmentation by ControlNet-added stable diffusion for multi-class weed detection. Computers and Electronics in Agriculture, 232, 110123.

[20] Cheng, H. Y., Su, C. C., Jiang, C. L., & Yu, C. C. (2025). Pose Transfer with Multi-Scale Features Combined with Latent Diffusion Model and ControlNet. Electronics, 14(6), 1179.

[21] Jääskeläinen, P., Sharma, N. K., Pallett, H., & Åsberg, C. (2025). Intersectional analysis of visual generative AI: the case of stable diffusion. AI & SOCIETY, 1-22.