Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

WDANet: exploring style feature via diffusion model for woodcut-style design

Yangchunxue Ou | Jingjun Xu

[1]Film-Video-Animation School, Sichuan Fine Arts Institute, Chongqing, China

Correspondence
Corresponding author Jingjun Xu, Sichuan Fine Arts Institute, 28 Chengnan Road, Shapingba District, Chongqing, China.
Email: 13368286806@163.com

Present address
Sichuan Fine Arts Institute, 28 Chengnan Road, Shapingba District, Chongqing, China.

Abstract

Due to the strong visual impact and color contrast inherent in woodcut style design, it has been applied in animation and comics. However, traditional woodcuts, hand-drawn, and previous computer-aided methods have yet to address the issues of dwindling design inspiration, lengthy production times, and complex adjustment procedures. We propose a novel network framework, the Woodcut-style Design Assistant Network (WDANet), to tackle these challenges. Notably, our research is the first to utilize diffusion models to streamline the woodcut-style design process. We curate the Woodcut-62 dataset, which features works from 62 renowned historical artists, to train WDANet in absorbing and learning the aesthetic nuances of woodcut prints, offering users a wealth of design references. Based on a denoising network, our WDANet effectively integrates text and woodcut-style image features. WDANet allows users to input or slightly modify a text description to quickly generate accurate, high-quality woodcut-style designs, saving time and offering flexibility. As confirmed by user studies, quantitative and qualitative analyses show that WDANet outperforms the current state-of-the-art in generating woodcut-style images and proves its value as a design aid.

KEYWORDS
woodcut-style design, diffusion model, computer-aided design, text-to-image model

1 | INTRODUCTION

Woodcut, as an ancient and enduring art form, holds a significant place in the history of visual expression. Originating as a method of printing reverse images onto wooden boards and transferring them onto paper, woodcut has long been revered for its simplicity and reproducibility. Historically, woodcut prints have been characterized by their stark black-and-white monochrome palette, offering artists a means of precise control over light and shadow[1]. The bold contrasts and textural quality of woodcut marks have endowed this style with a unique visual impact, captivating viewers across generations.

In the realm of animation, the influence of woodcut-style design is palpable, offering a distinctive aesthetic for visual storytelling and comic art. However, replicating the intricate texture of woodcut marks through traditional hand-drawn methods presents a formidable challenge to animators. In response, computer-aided techniques have emerged, aiming to simulate the organic quality of woodcut while streamlining the production process. This intersection of artistry and technology has given rise to a burgeoning field of research in Non-Photorealistic Rendering (NPR)[2], with a focus on translating the essence of woodcut into digital form. Despite significant advancements in simulating woodcut texture, existing techniques have limitations. They often prioritize visual fidelity over artistic expression, lacking the nuanced understanding of composition, thematic elements, and stylistic nuances inherent in woodcut design. Moreover, the preparatory work required to generate high-quality inputs for these algorithms can be labor-intensive, hindering the creative process.

Recently, large-scale models in the realm of generative art, such as Stable Diffusion (SD)[3], Midjourney[4], DALL-E 3[5], have indeed provided new opportunities for artists. Despite this, the typical method to control the model and generate specific designs involves fine-tuning, which entails a substantial number of parameters. Some more efficient techniques include freezing the basic model parameters and exclusively training additional modules[6–8]. Research datasets available can drive advancements

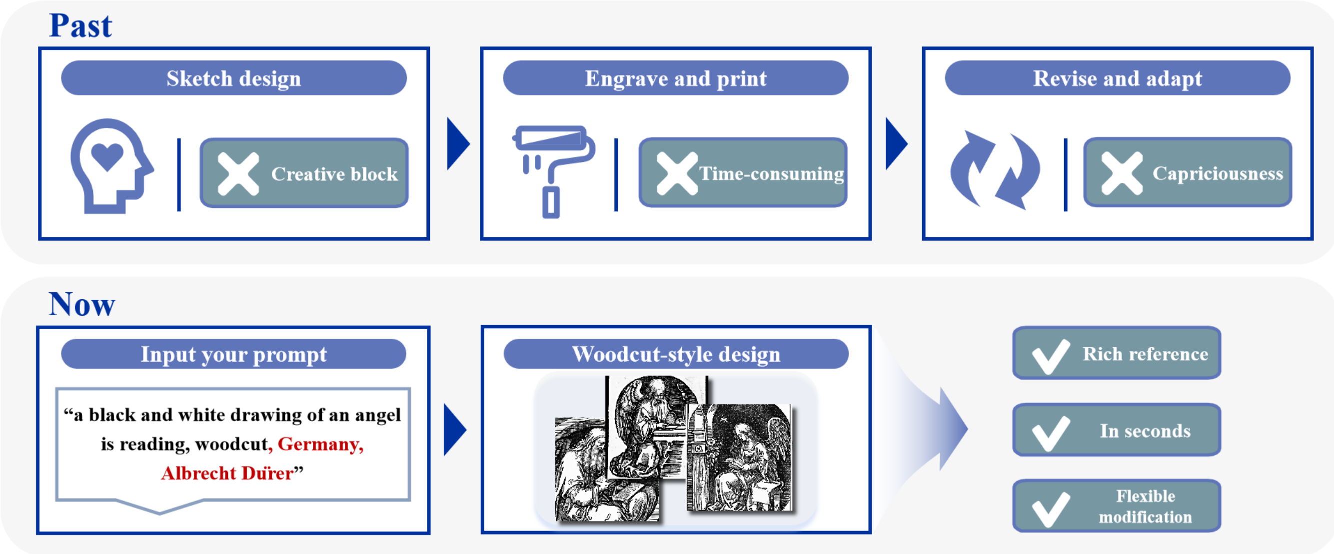

FIGURE 1 Considering the significant challenges encountered in previous stages of woodcut-style design, WDANet simplifies the process by enabling users to input and modify text descriptions. In seconds, WDANet generates diverse woodcut images tailored to meet user specifications. The text highlighted in red indicates that WDANet supports replicating specific artistic styles or employing generic styles.

within a field. For example, the introduction of the VLD 1.0 dataset[9] significantly propelled log detection and recognition. Nevertheless, the domain of woodcut-style design remains largely unexplored, and a high-quality woodcut dataset is needed, along with the absence of a standardized processing pipeline.

Therefore, this paper introduces a dataset with images and textual annotations named Woodcut-62. This dataset is the dataset tailored explicitly for the woodcut style. Building upon this foundation, we present a groundbreaking model called Woodcut-style Design Assistant Network(WDANet), which, with lightweight training parameters, guides textual inputs to generate design references meeting specified criteria. Compared with previous methods, the advantages of our proposed method are shown in Fig. 1

In summary, this paper outlines the contributions as follows:

- We curate Woodcut-62, a high-quality dataset tailored for woodcut-style designs, comprising 3,058 text-image pairs. This dataset offers style references from 62 artists representing eight countries who greatly influenced history.
- Introducing the lightweight architecture WDANet, we pioneer the application of diffusion models in a woodcut-style design. WDANet accurately emulates the style of a specific artist and generates woodcut-style designs within seconds, relying solely on input text.
- The fidelity, diversity, and aesthetic characteristics of the woodcut style design drawings generated with WDANet guidance have achieved SOTA while garnering better user preferences in the comprehensive artistic evaluation conducted during the user study.

2 | RELATED WORK

2.1 | Woodcut-style design

The pioneering application of computer graphics in woodcut design is initiated by Mizuno et al.[10]. They integrate traditional woodcut production into an interactive simulation system, allowing users to carve within a virtual 3D space. This system synthesizes woodcut images with authentic printing effects on a 2D grid of virtual paper. Subsequent works by Mizuno et al.[11,12] perfect features such as automatic engraving based on grayscale map characteristics and color printing. Their research mainly focuses on the Ukiyo-e style[13,14]. To enhance the user experience, they were pioneers in migrating the system to a pressure-sensitive pen and tablet. Mello et al.[15] are the first to introduce image-based artistic rendering, specifically generating woodcut-style images. They obtain rendered images through image segmentation, direction field computation, and stroke generation. However, the simulated woodcut scores are relatively basic. Building upon this work, Jie Li et al.[16,17] allocate fractions gathered from authentic woodcut textures according to segmented areas, applying this method to Yunnan out-of-print woodcut. A recent study by Mesquita et al.[18] proposes woodcut generation based on reaction-diffusion, enhancing woodcut representation and user control by introducing noise.

Current research in woodcut-style image generation primarily focuses on simulating woodcut marks, sidelining the artistic design aspect of woodcut. Achieving aesthetically pleasing picture arrangements still requires user intervention. Woodcut-style design generation relies on rendering based on 3D models or images, where the quality and depth of prerequisites significantly impact the output. Crafting a visually appealing woodcut artwork entails a high threshold and considerable time investment. Additionally, the existing algorithms concentrate on specific styles, such as Ukiyo-e and Yunnan out-of-print woodcuts, lacking adaptability to versatile artistic styles.

2.2 | Computer-aided art generation

Over the past few decades, a series of rendering and texture synthesis algorithms proposed by computer-aided design research[19–21] have predominantly focused on image stylization. The intervention of artificial intelligence in art introduces more creative applications[22,23]. The confluence of these fields can be traced back to the advent of Generative Adversarial Networks (GAN)[24], where two neural networks—generators and discriminators—compete to produce images resembling actual sample distributions. DeepDream[25] is the first to explore neural networks' potential to inspire artistic creation. Utilizing convolutional neural networks (CNNs), DeepDream transforms input images into highly interpretable, dream-like visualizations. Another CNN-based study, A Neural Algorithm of Artistic Style[26], separates and recombines semantics and style in natural images, pioneering Neural Style Transfer (NST). Although NST significantly influences artistic style treatment, it typically learns stylistic traits from existing images rather than facilitating original artistic creation. Elgammal et al. introduce Creative Adversarial Networks (CAN)[27] based on GAN, aiming to maximize divergence from established styles while preserving artistic distribution and fostering creativity.

The emergence of Transformer[28,29] propels the application of multimodal neural networks notably in many tasks[30,31]. For example, Hu et al.[32] proposed utilizing transformers for roof extraction and height estimation. Similarly, in the field of remote sensing detection[33,34] and re-identification[35–37], transformers have demonstrated tremendous potential. Recently, text-guided image generation through generative artificial intelligence, such as Glide[38], Cogview[39], Imagen[40], Make-a-scene[41], ediffi[42] and Raphael[43] gain widespread use, particularly with significant advancements in large-scale diffusion models[44]. In order to achieve favorable results on the downstream tasks, the past practice is to spend substantial technical resources to fine-tune the model. Many existing studies build upon SD, incorporating adapters for guided generation and requiring minimal additional training while keeping the original model parameters frozen. ControlNet[6] pioneered this method to learn specific task input criteria, such as Depth Map[45], Canny Edge[46]. ControlNet Reference-only efficiently transfers the style and subject from a reference diagram while conforming to textual descriptions, eliminating the need for additional training. In the T2I-Adapter[7] approach, the reference image is fed into the style adapter to extract stylistic features and integrate them with text features. Uni-ControlNet[47] employs two adapters—one for global and one for local control—enabling the combinability of diverse conditions. The IP-Adapter[8]supports the fusion of individual image features with text features via the cross-attention layer, ensuring the preservation of both the main subject of the image and its stylistic attributes. Additionally, PCDMs[48] suggest employing inpainting to fine-tune the entire SD for achieving pose-guided human generation. Nonetheless, this approach of training by fully releasing all parameters is not practically economical.

While these adapters effectively generate similar styles from image-guided models, the generated results tend to closely resemble specific reference images, posing challenges in providing valuable references within the design realm. Artistic creation often requires diverse images to draw inspiration from when exploring various themes. Therefore, generated images must exhibit diversity while maintaining coherence between visuals and accompanying text descriptions. Our study centers on woodcut-style design, employing dual cross-attention mechanisms to harmonize visual and textual features. Moreover, our approach facilitates the transfer of a specific artist's style across different thematic contexts.

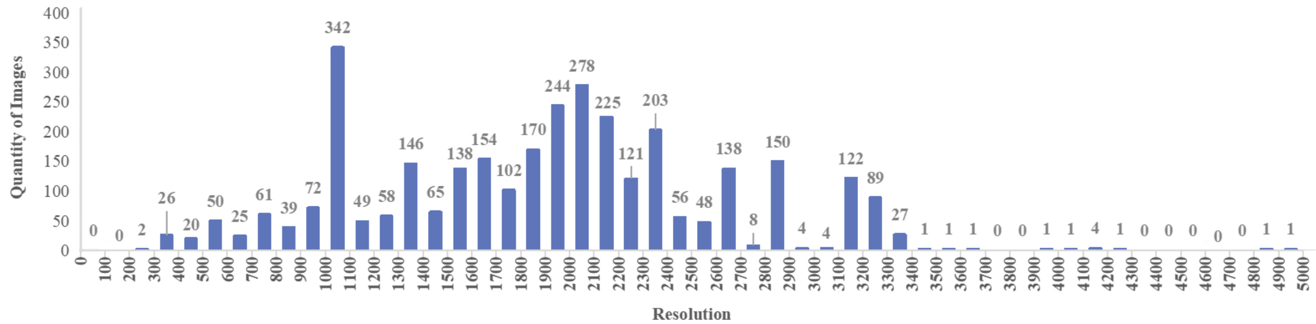

FIGURE 2 The resolution details of the 3,248 collected original images, categorized by the shortest edge of each image, illustrate that most resolutions are concentrated within 500 to 3400 pixels. Notably, a significant portion of these images surpasses the 512 × 512 pixel mark.

3 | METHOD

3.1 | Woodcut-62

3.1.1 | Collect

Numerous datasets of artworks cater to a wide range of machine learning tasks, such as VisualLink[49], Art500k[50], and ArtEmis[51]. However, the domain of woodcut prints still needs to be explored regarding available datasets. We gather 3,248 open-access woodcut images from Web Gallery of Art[†], National Gallery of Art[‡], and Wikimedia Commons[§]. The distribution of resolutions is depicted in Fig. 2. We manually engage art experts to adjust these artworks to 512 × 512 resolution to standardize resolution and minimize training costs. Tailoring involves preserving the thematic essence and the most aesthetic characteristics. Larger-sized images are split into multiple images for better handling. As woodcuts are prints, their electronic copies lose some detail compared to the originals. Working closely with experts, we manually adjusted image parameters to enhance features, maintaining the characteristic texture while reducing noise. Under art expert guidance, we filter out low-resolution or duplicated images, resulting in 3058 high-quality woodcut images. These images represent 62 historically influential artists from eight countries, spanning genres like romanticism, abstract expressionism, and realism. Ultimately, this effort culminated in creating the inaugural woodcut style dataset for the text-to-image task—Woodcut-62. The style distribution of the dataset is shown in Fig. 3.

3.1.2 | Label

To accommodate the text-to-image generation task, we transition from the original caption to using natural language for annotation. This label process is outlined in Fig. 4. Initially, we experiment with the BLIP[52]Captioning and Filtering (CapFilt), a new dataset bootstrapping method that filters out noisy titles and automatically composites captions for images. However, upon manual inspection, we discovered that this method's efficacy could be improved, particularly for abstract expressionism or intricate content. To address this, we seek the expertise of art specialists to fine-tune annotations for these specific types of images manually. To distinctly mark the woodcut style, we append the keyword "woodcut" to each image, along with details such as the artist's name and the country of origin. This meticulous labeling ensures that the model can discern between various artistic styles. Finally, we encapsulate text-image pair labels into a JSON file.

† www.wga.hu
‡ www.nga.gov
§ https://commons.wikimedia.org

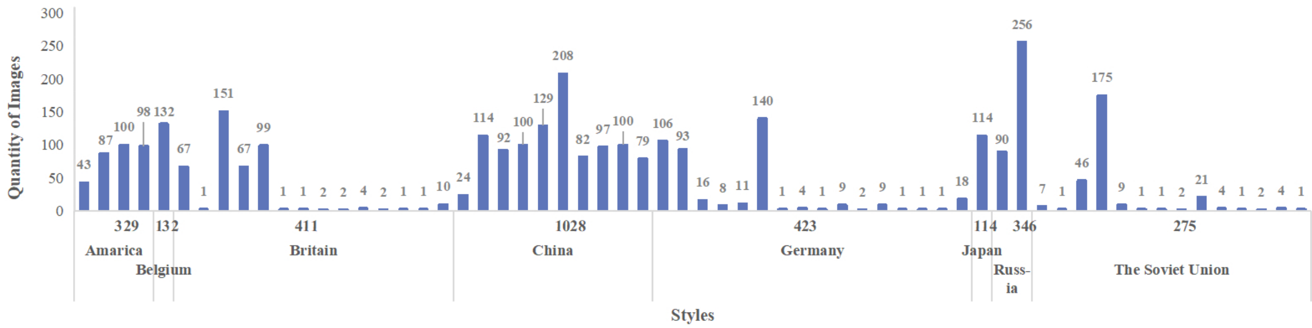

FIGURE 3 Woodcut-62 is categorized based on the styles of 62 historically influential Woodcut artists, encompassing artworks from eight distinct countries, reflecting a diverse style distribution.

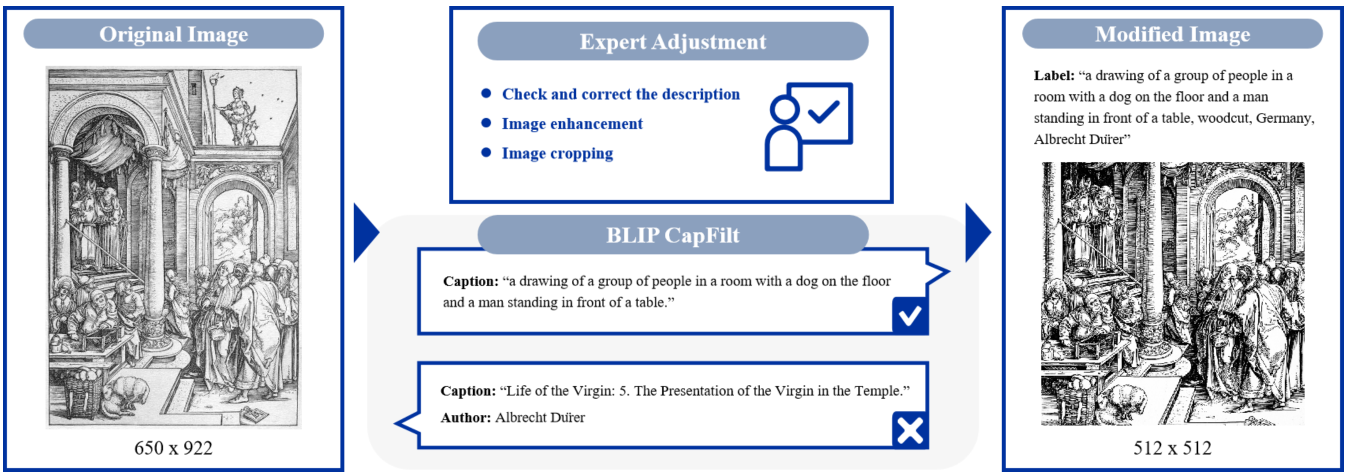

FIGURE 4 The image processing workflow in Woodcut-62 involves several steps. Initially, the collected raw images undergo captioning by BLIP CapFilt, resulting in adaptive captions. With guidance from art experts, we meticulously verify and rectify the descriptions. Additionally, manual cropping and enhancement techniques are applied to the original images, resulting in uniformly sized 512 $\times$ 512 images and corresponding labels.

3.2 | Prelimiaries

Diffusion model[53] is an important method for describing complex systems' dynamic behavior and stable state distribution. The image generation is likened to the diffusion of ink in water. It is divided into two processes. The forward diffusion process transforms the initial data Z_0 into a Gaussian distribution Z_T by iteratively adding random Gaussian noise through T iterations. The reverse denoising process involves predicting noise ϵ_θ from Z_T and gradually restoring Z_0. When $t \in [0, T]$, the noise intensity is β_t, if $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, then Z_t is the linear combination of the original signal Z_0 and random noise ϵ, meeting:

$$Z_t = \sqrt{\bar{\alpha}_t} Z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon. \tag{1}$$

Governed by condition c, the simplified variant loss function of the training model is as follows:

$$L_{simple} = \mathbb{E}_{Z_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), c, t} \|\epsilon - \epsilon_\theta(Z_t, c, t)\|^2. \tag{2}$$

Where $\mathbf{I}$ represents the identity matrix.

Balancing sample fidelity and pattern coverage in conditional diffusion models commonly involves employing a classifier as guidance, necessitating the additional training of an image classifier separate from the diffusion model. However, classifier-free guidance, as proposed in[54], trains both conditional and unconditional diffusion models, randomly discarding the control conditions c. The noise prediction during the sampling stage adheres to the following formula:

$$\widetilde{\epsilon}_\theta\left(Z_t, c, t\right) = \omega \epsilon_\theta\left(Z_t, c, t\right) + (1-\omega)\epsilon_\theta\left(Z_t, t\right). \tag{3}$$

Here, ω represents the scale guiding the faithfulness to the control conditions c in the generated results. In this study, the foundational generative model utilized is SD[3]. SD maps the image to the low-dimensional latent space instead of the pixel space for diffusion, significantly reducing the diffusion model's computation. Within SD, a U-Net[55] is trained as the core network for the noise prediction model. A cross-attention mechanism is introduced to integrate control information c into the intermediate layers of the U-Net. This inclusion forms a conditional denoising autoencoder, effectively governing the image synthesis. The cross-attention layer yields the following output:

$$\begin{cases} Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \\ Q = W_Q \cdot \varphi(Z_t), \quad K = W_K \cdot \tau_\theta(c), \quad V = W_V \cdot \tau_\theta(c). \end{cases} \tag{4}$$

Where W_Q, W_K, W_V are trainable weight parameter matrices, $\varphi(Z_t)$ represents a learnable flattened embedding of the U-Net, and $\tau_\theta(c)$ denotes a learnable domain-specific encoder that converts c into an intermediate representation.

3.3 | WDANet

It is critical to preserve the thematic essence described in the text for design image generation while learning coarse-grained details like style and composition from woodcut images. Previous methodologies often involved superimposing text and other features via a cross-attention layer in the diffusion model, which could be more effective in transferring and balancing various control conditions. We enhance the control of woodcut style and texture based on the work of decoupled cross-attention[8].

In the text-to-image task, CLIP[56] plays a crucial role as the link between text and image. CLIP comprises a text encoder and an image encoder. The text and image modalities could be aligned in the feature space through contrastive learning and extensive training on numerous text-image pairs. To incorporate Woodcut-style guidance into the text-to-image diffusion model, we propose an SD-based frozen CLIP image encoder. Additionally, we add a trainable linear layer (Linear) and a Layer Normalization[57](LayNorm). Their goal is to extract the scattered woodcut image features x and map them to the feature c' that best matches the target effect and aligns with the text embedding dimension. The Linear obtains $y = W'x + b$, and the LayNorm transforms the input into a control condition with woodcut style characteristics c':

$$\begin{cases} c' = g \odot \frac{(y-\mu)}{\sqrt{\sigma^2+\varepsilon}} + b', \\ \mu = \frac{1}{H}\sum_{i=1}^{H} y_i, \quad \sigma = \sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(y_i-\mu\right)^2}. \end{cases} \tag{5}$$

In the above formula, W', b, g, b' are learnable parameters, H is the number of hidden units, and $\odot$ is the element-wise multiplication. Next, we can add c' to the new cross-attention layer:

$$\begin{cases} Attention(Q, K', V') = softmax(\frac{QK'^T}{\sqrt{d}}) \cdot V, \\ Q = W_Q \cdot \varphi(Z_t), \quad K' = W'_K \cdot \tau_\theta(c'), \quad V' = W'_V \cdot \tau_\theta(c'). \end{cases} \tag{6}$$

Where W_K and W_V are learnable parameters for the new cross-attention layer. We introduce α (default $\alpha = 0.5$) to balance the fidelity and diversity of the generated images by controlling weights. When $\alpha = 0$, WDANet reverted to the initial text-to-image model (i.e., SD). In this way, the cross-attention layers with text and woodcut style characteristics, respectively, can be combined into dual cross-attention:

$$Attention_{\text{new}} = (1-\alpha)\text{Attention}(Q, K, V) + \alpha\text{Attention}(Q, K', V'). \tag{7}$$

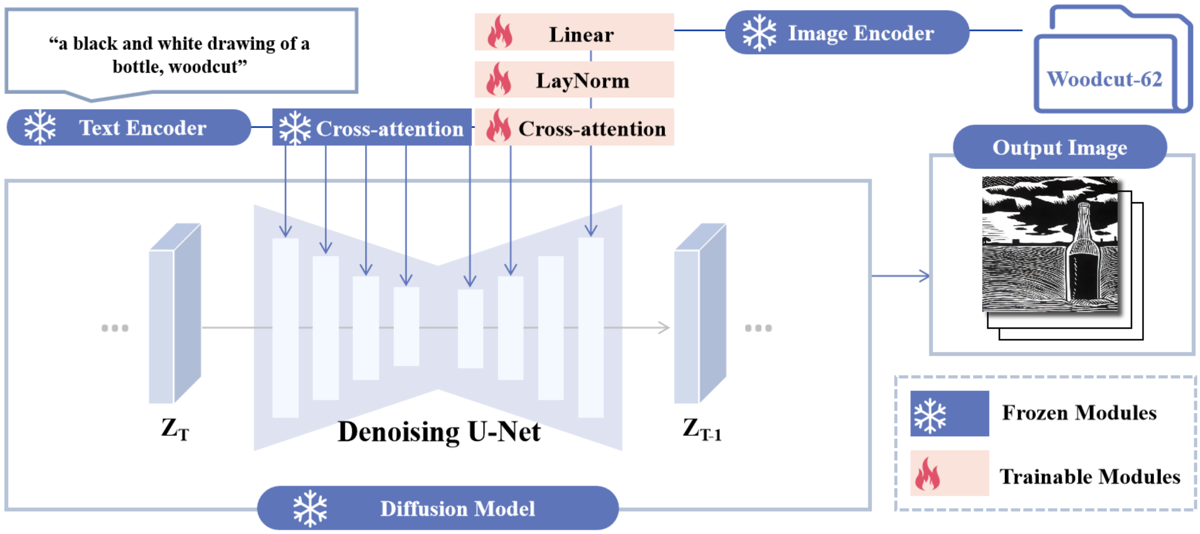

FIGURE 5 The overall framework of WDANet. WDANet only trains Linear, LayNorm, and an added cross-attention layer while freezing the parameters of the other modules. The user input prompt aligns with the woodcut-style image embedded in Woodcut-62, guiding the diffusion model via dual cross-attention to generate woodcut-style designs.

The structure of WDANet is depicted in Fig. 5. Woodcut-62 served as the inherent input for the image encoder, eliminating the need for users to provide a specific Woodcut reference image. Instead, inputting text enabled WDANet to query Woodcut-62 to match either the artist's style or a universal woodcut aesthetic. While seemingly straightforward, this approach relied more on text for WDANet's style control than a single image, preventing an excessive focus on granularity that might compromise generative diversity.

3.4 | Training and inference strategies

In SD, the text embedding is derived from the input text using the pre-trained CLIP text encoder, which serves as the guiding condition for the denoising process. However, relying solely on text guidance is insufficient for the model to grasp the aesthetic features of woodblock prints. To address this problem, the added image encoder provided finer details, yet fine-tuning parameters remained necessary to adapt it to the woodcut design task. Following the methodology of training the Adapter with freezing CLIP and SD, our proposed WDANet only needs to learn parameters for Linear, LayNorm, and a cross-attention layer to achieve promising outcomes.

In the training stage, we randomly take the images from the dataset Woodcut-62 and the matching labels as inputs, map the woodcut-style features to the same dimension as the text features and inject the joint features into each middle layer of U-Net (16 layers in total) through dual cross-attention. Although there are some approaches to fuse features using parallel attention mechanisms[58], ensuring the inheritance of diversity from the original model while enabling the newly added cross-attention to align with the woodcut-style theme is crucial. Inspired by using a unique identifier to fine-tune in Dreambooth[59], we embed the keyword "woodcut" along with additional country and artist information into the semantic priors during training. This approach allows WDANet to encompass abstract and extended semantic features, such as textual cues, and accurate and concrete visual information from woodcut images. We borrow the concept of classifier-free guidance and set the image conditions or text conditions as empty according to the probability of $p_{\varnothing} = 0.05$ and set both conditions as empty at the same time under the same probability, that is $p(c = \varnothing), p(c', \varnothing), p(c = \varnothing \cap c' = \varnothing) = 0.05$. The training objectives we utilized are:

$$L_{new} = \mathbb{E}_{Z_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), c, c', t} \left\| \epsilon - \epsilon_\theta \left(Z_t, c, c', t \right) \right\|^2 . \tag{8}$$

Accordingly, the target of noise prediction is:

$$\widetilde{\epsilon}_\theta\left(Z_t, c, c', t\right) = \omega\epsilon_\theta\left(Z_t, c, c', t\right) + (1-\omega)\epsilon_\theta\left(Z_t, t\right). \tag{9}$$

During the inference process, we conceptualize the multi-condition diffusion model as a text-to-image framework. On one side, the image encoder is linked to the Woodcut-62 dataset to enhance the Woodcut style guidance. Meanwhile, the user has access only to the text input port. In comparison to the original SD approach, this modification strengthens the visual features of the concrete and simplifies the overall process when compared to the multi-condition diffusion model. To expedite the generation of woodcut-style designs, we employed DDIM[60], an accelerated sampling method utilizing a non-Markov diffusion process to simulate the reverse process of Markov diffusion.

4 | EXPERIMENT

4.1 | Implementation details

We utilize CLIP ViT-H/14[61] as the image encoder and SD v1.5 [¶] as the backbone network. WDANet's parameters were lightweight and enabled training on a 24GB RTX 3090 graphics card with 80GB RAM. Employing AdamW[62] as the optimizer, we train for 70k global steps on Woodcut-62 with a learning rate set at 1×10^{-4} and a weight decay of 0.01. DDIM served as the sampling acceleration algorithm, executing 50 sampling steps during inference. The coefficient α in dual cross-attention defaulted to 0.5. The prompt on the text input side should include the keyword "woodcut" to identify the woodcut style, and if the generation task needs to be specific to the artist's style, then follow the country and artist name, separated by commas.

4.2 | Quantitative evaluation

Our aim in assisting design is to ensure that the model-generated woodblock prints exhibit diverse compositions, a rich array of elements, and aesthetically pleasing arrangements based on a given thematic description. Therefore, our chosen metrics primarily assess the alignment between text and image and the aesthetic quality of the generated images. We utilize CLIPScore[63] based on CLIP ViT-B/32 to assess the alignment between text and image. As a supplementary evaluation criterion, BLIP[52] forecasts fine-grained alignment between visual and textual elements via a linear layer. Its output includes the likelihood of image-text matching (ITM) and the cosine similarity between the features of both modalities. When evaluating the influence of generated images on users' aesthetic preferences, we incorporate ImageReward[64], a model trained on a dataset of 137k expert comparisons. This model automatically gauges the quality of text-to-image conversions and closely aligns with human preferences.

TABLE 1 Our proposed WDANet and SOTA methods assess the alignment between text and image and the quantitative evaluation of aesthetic preferences, with SD as the baseline.

Method	**CLIP Score↑**	**BLIP-ITM↑**	**BLIP-Cosine↑**	**ImageReward↑**
Baseline[3]	18.52	76.42%	0.4077	-2.08
ControlNet[6]	27.11	98.97%	0.4192	-0.08
T2I-Adapter (Style)[7]	28.03	94.53%	0.4693	-0.95
IP-Adapter[8]	29.73	99.81%	0.4776	0.15
Ours	**34.20**	**99.98%**	**0.4826**	**0.66**

We present a comparative analysis between our approach, WDANet, and several state-of-the-art methods such as ControlNet, T2I-Adapter, and IP-Adapter. We provide empirical data to showcase the advancements over the original SD model. To ensure

[¶] https://huggingface.co/runwayml/stable-diffusion-v1-5 diverse coverage of text descriptions across different categories and difficulties, we randomly select ten prompts from the PartiPrompts[65], which includes over 1600 English prompts. We generate 100 woodcut-style images from each prompt in the format ”a black and white drawing of *, woodcut.” [#]The ultimate score is calculated by averaging the results from each dataset. The results in Table 1 indicate that our proposed WDANet achieved a CLIP Score of 34.20, nearly doubling the baseline score of 18.52. The BLIP-ITM (the likelihood of image-text matching) reaches 99.98%, confirming a high degree of alignment between the generated images by WDANet and the input text. Moreover, the BLIP-Cosine similarity between text and image features is notably high, recorded at 0.4826. When evaluating aesthetic preferences using ImageReward, our approach exhibits a 2.74 improvement over the baseline, surpassing the second-ranking IP-Adapter by 0.51. WDANet achieves the SOTA level in generating woodcut-style images.

4.3 | Qualitative analysis

In this study segment, we present woodcut-style images generated by prompts extracted from the PartiPrompts corpus using various methods. SD serves as the baseline for the text-to-image model, while other models modulate style features by incorporating an additional woodcut image.

In Fig. 6, we observe that SD (baseline) and ControlNet display instability, while T2I-Adapter and IP-Adapter tend to mirror the composition, characters, or actions of the reference image too closely, as seen in figure (1), where both T2I-Adapter and IP-Adapter depict middle-aged men with a beard. Our method, however, focuses on learning the stylistic and textural features of woodcut images. In (5), WDANet showcases a more accurate representation of complex prompts such as ”a woman with long hair” and ”a luminescent bird,” illustrating them in an incredibly artistic and expressive manner.

In Fig. 7, we standardize the input to the same reference image to examine the diverse outputs generated by various methods under different prompts. Baseline and ControlNet struggle to adopt the woodcut style. However, when examining the columns for the outputs of T2I-Adapter and IP-Adapter separately, they exhibit striking similarities across different prompts, evident in the figure, where IP-Adapter consistently depicts a towering tree on the left side. In contrast, our method demonstrates strong performance in both diversity and fidelity. Notably, when presented with long and complex prompts, as depicted in (4), WDANet accurately generates woodcut-style designs that align with the description.

4.4 | User study

User study evaluates four aspects: visual-textual consistency, learning style precision, aesthetic preference, and generative diversity. Users are asked to select the most effective image among those generated by the four models. Visual-textual consistency measures the alignment between the model-generated image and the provided text, akin to assessing a designer’s ability to create thematic designs without deviating from the subject—a fundamental requirement. Learning style precision assesses the model’s capability to replicate an artist’s style. As the artist Walter Darby Bannard expressed, ”When inspiration dies, imitation thrives,” suggesting that designers often enhance their skills by imitating masterpieces. Aesthetic preference gauges the overall quality of the generated woodcut images based on adherence to aesthetic principles like composition, element arrangement, and visual harmony. Generative diversity tests revealed that models with additional image-controlling conditions are often constrained by single-image features, leading to a loss of richness in the generated images. Consequently, concerning this aspect, we directly compare WDANet to state-of-the-art text-to-image models, Midjourney and DALL-E 3.

We collected 68 questionnaires from 14 cities in China, with 38 respondents having received art education and possessing some knowledge of woodcuts. In the questionnaire presented to users, each option is arranged randomly. For comparison purposes, in these cases, we have fixed the options. In our questionnaire Fig. 8, *A* represents ControlNet, *B* represents the image generated by T2I-Adapter, *C* depicts IP-Adapter’s output, and *D* represents WDANet, proposed by us. In *Case* 1, our generated results align best with the highlighted keywords. In contrast, option *A* lacks the imitation of woodcut style, and the actions portrayed in *B* and *C* are not sufficiently accurate. For *Case* 2, we evaluate the Learning style precision of different methods using Kathe Kollwitz’s woodcut works as reference. As a German expressionist printmaker, her pieces evoke a heavy and melancholic atmosphere, where our approach closely aligns with her characterization and thematic tone. In *Case* 3, option *D* exhibits a stable effect while comparing Aesthetic preference, illustrating ”a retro town by the river” with a tranquil aesthetic.

[#]”*” Indicates replaceable prompt. The text input in this format is used to generate images that are closer to a woodcut than a realistic photograph.

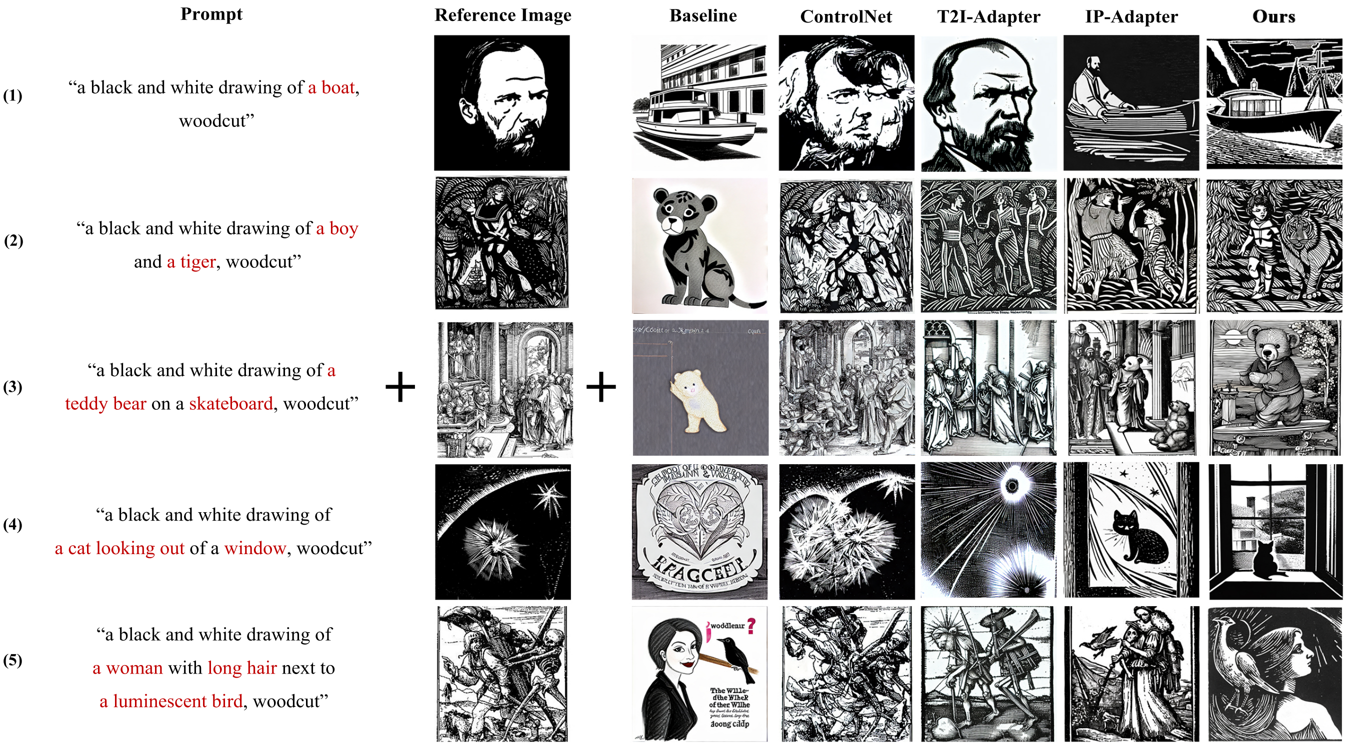

FIGURE 6 The visual comparison result between WDANet and other methods under specific conditions. The prompt is randomly selected from the PartiPrompts corpus, with the reference image embedded as a control for woodcut style conditions. Those highlighted in red are keywords.

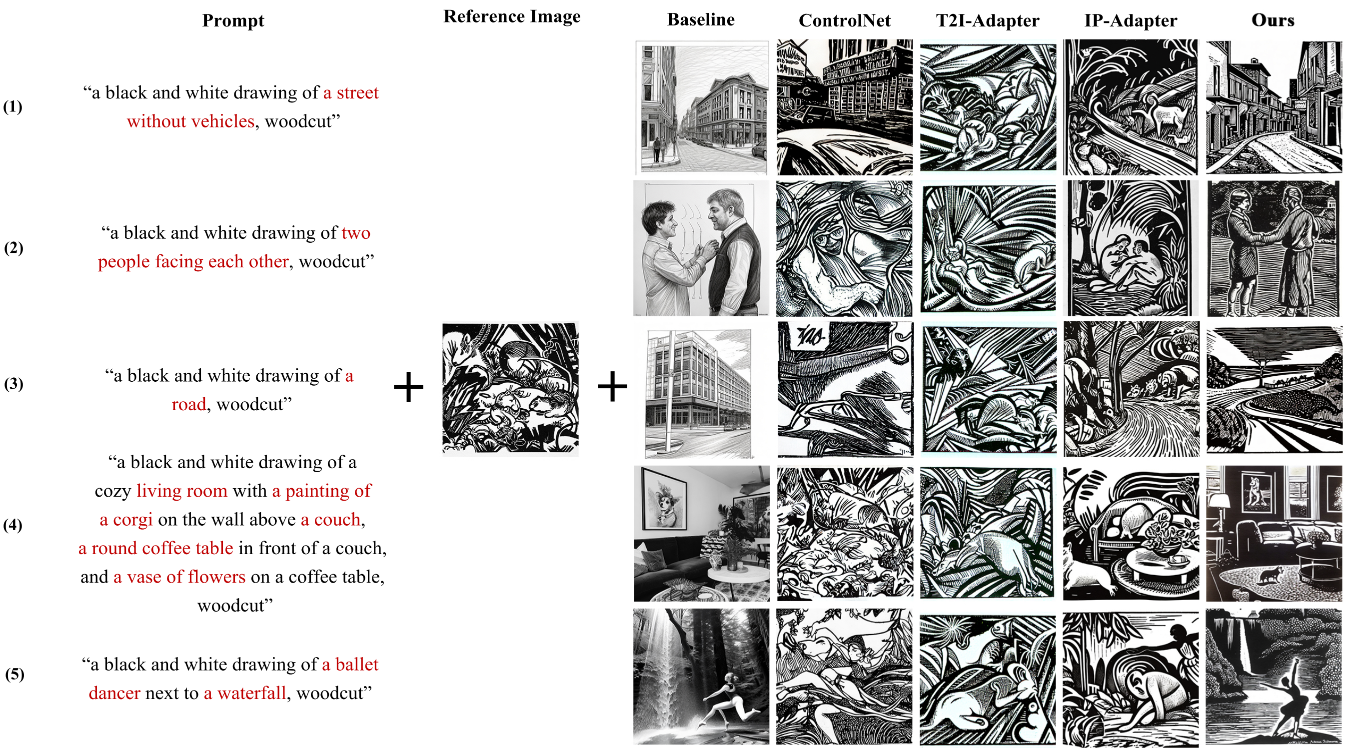

FIGURE 7 Given the same reference image to control variables and inputting different prompts, our method is compared with other models, primarily to examine the diversity of the generated images.

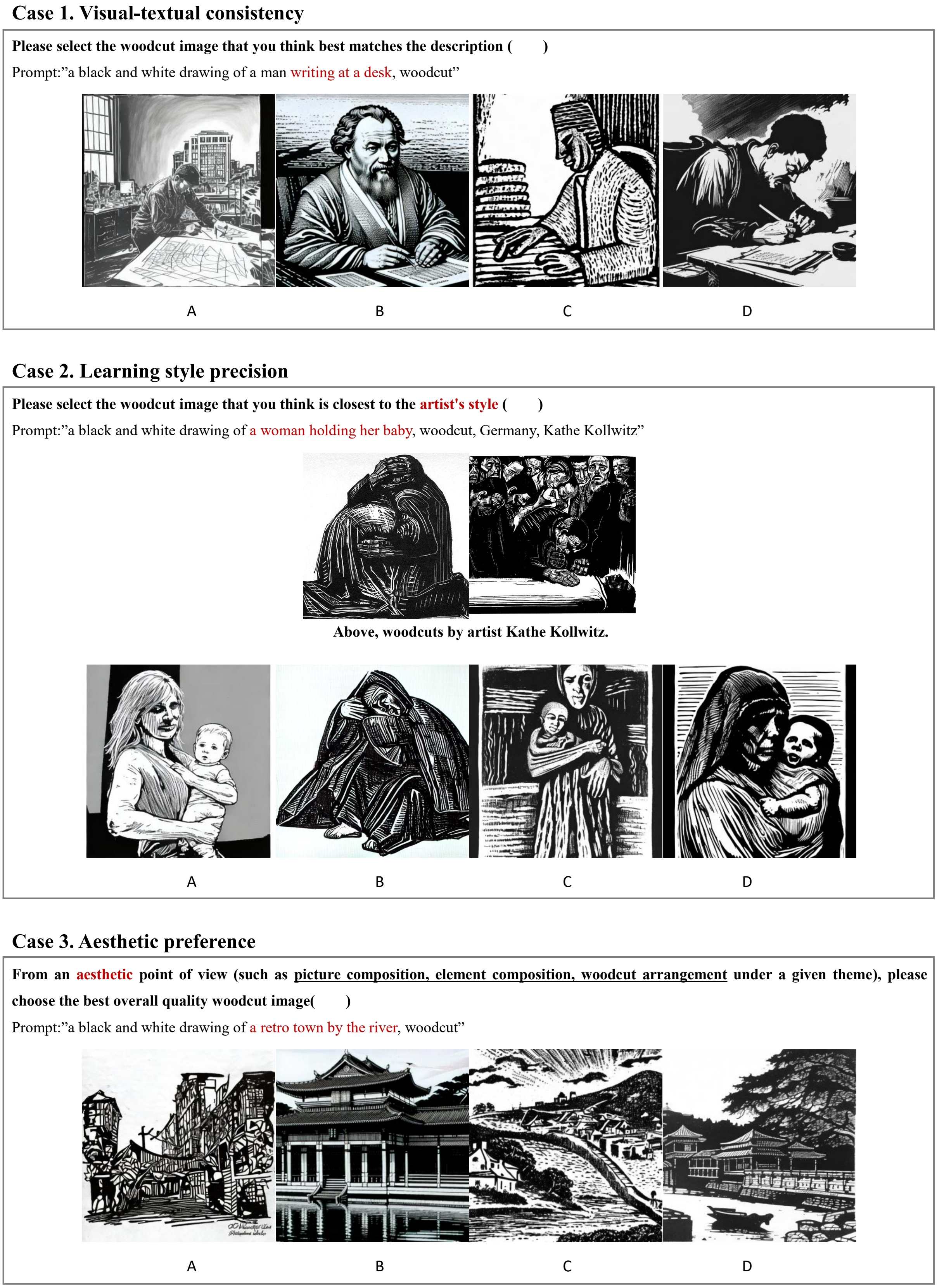

FIGURE 8 We present questionnaire cases evaluating visual-textual consistency, learning style precision, and aesthetic preference. In these cases, *A* represents ControlNet, *B* stands for T2I-Adapter, *C* depicts IP-Adapter, and *D* represents WDANet, our proposed method.

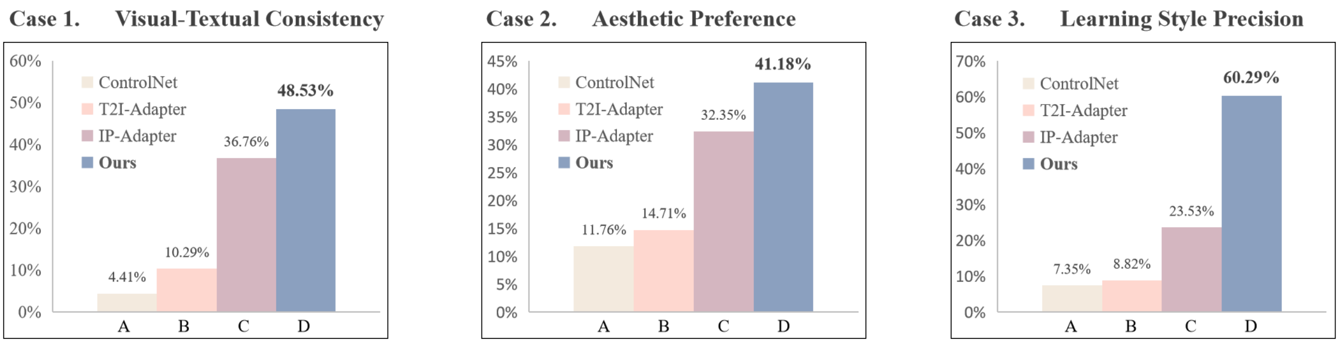

FIGURE 9 The user study results comparing our WDANet with SOTA methods indicate user preferences. Here introduce three metrics for evaluating the generative model for design tasks: visual-textual consistency, learning style precision, and aesthetic preference. The corresponding questionnaire content can be seen in Fig. 8.

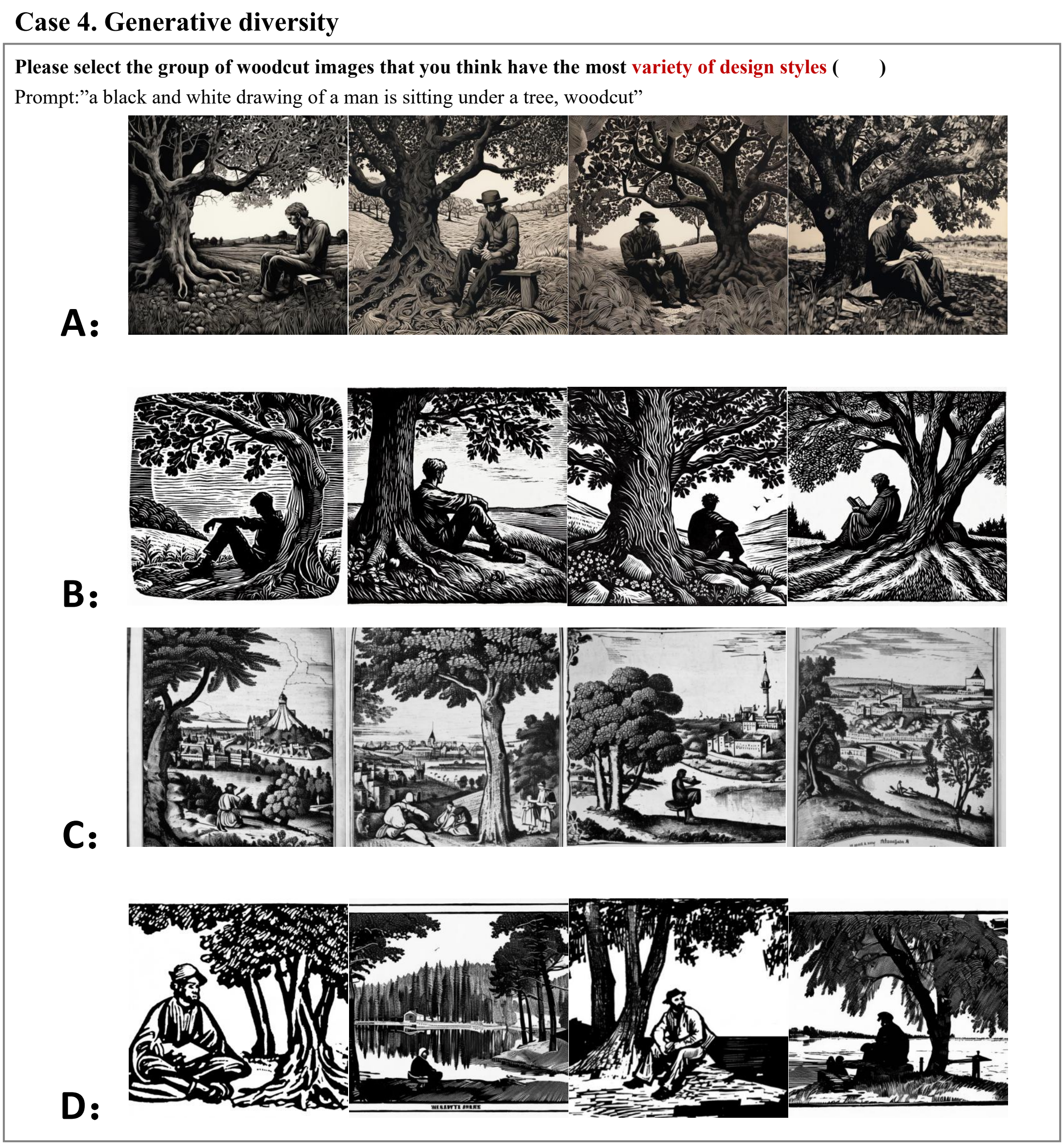

FIGURE 10 Our method is compared against state-of-the-art text-to-image models and adapter methods to assess generative diversity. *A*, *B*, *C*, and *D* correspond to Midjourney, DALL-E 3, IP-Adapter, and WDANet, respectively.

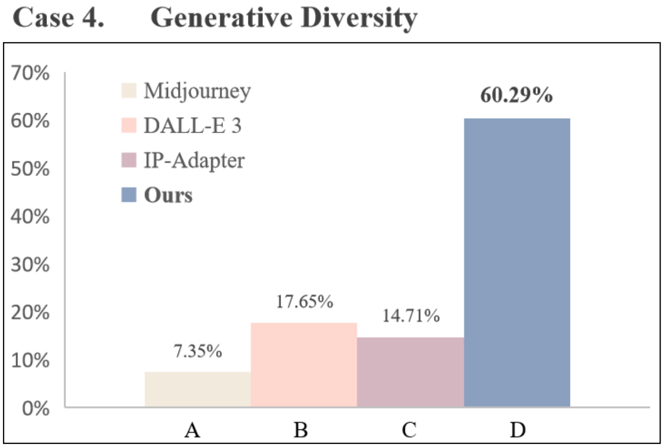

FIGURE 11 In *Case* 4, the questionnaire results regarding generative diversity showcased that WDANet demonstrated a notable advantage in design richness, securing 60.29% of the votes.

Correspondingly, the collected data shown in Fig. 9 demonstrates that WDANet gains more user preference across all three aspects of the evaluation.

In questionnaire *Case* 4, as Fig. 10 shows, we investigate the performance of different models in generating different designs, specifically regarding generative diversity. *A* represents Midjourney, *B* stands for DALL-E 3, *C* depicts IP-Adapter, which performs better among the adapter methods, and *D* represents WDANet, our proposed method. It is noticeable that options *A* and *B* favor realism over artistic rendering in generating woodcut-style designs. Additionally, the diversity in picture composition, element arrangement, and texture distribution in options *A*, *B*, and *C* appears relatively limited. Conversely, our method demonstrates both stylistic variability and stable results. As the corresponding data in Fig. 11, 60.29% of users chose *D*, affirming the diversity of our generated results.

5 | CONCLUSION AND LIMITATION

This Work introduces Woodcut-62, the dataset categorizing woodcut prints by artist styles, and proposes WDANet as the inaugural text-to-image diffusion network architecture for aiding woodcut-style design. WDANet unites text features with embedded woodcut style image characteristics, guiding the diffusion model to generate woodcut-style design references aligned with design requisites. Compared to conventional computer-assisted approaches, WDANet offers a broader spectrum of inspirational references, is less time-consuming, and offers greater adjustability. Quantitative and qualitative experiments demonstrate WDANet's superior performance in woodcut-style design compared to other conditionally controlled adapters. In the user study, WDANet outperformed in visual-textual consistency, learning style precision, aesthetic preference, and generative diversity—four indicators related to aesthetic design dimensions—garnering higher user preferences.

Despite the promising performance of WDANet in assisting woodcut-style design, its lightweight structure still relies on guided image embedding. While this framework has the potential to be extended to art categories beyond woodcut prints, it still necessitates fine-tuning. When applied to animation, games, and other downstream tasks, WDANet is better suited for a single frame, such as layout or scene design, and cannot ensure a specific element's continuous and stable appearance on the screen. Moving forward, we aim to explore more flexible methods for learning style features. Our goal is to accomplish design assistance tasks seamlessly without any need for input images, thereby enhancing the adaptability and autonomy of the system.

REFERENCES

1. Saff D, Sacilotto D. Printmaking: History and Process. Rinehart and Winston: Holt; 1978. ISBN 978-0-03-085663-1.
2. Hertzmann A. Introduction to 3d non-photorealistic rendering: Silhouettes and outlines. Non-Photorealistic Rendering SIGGRAPH. 1999;99(1).
3. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 10684–10695.
4. Borji A. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. arXiv preprint arXiv:221000586. 2022;.

5. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. In: International Conference on Machine Learning. PMLR; 2021. p. 8821–8831.
6. Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 3836–3847.
7. Mou C, Wang X, Xie L, Zhang J, Qi Z, Shan Y, et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:230208453. 2023;.
8. Ye H, Zhang J, Liu S, Han X, Yang W. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:230806721. 2023;.
9. Liu J, Shen F, Wei M, Zhang Y, Zeng H, Zhu J, et al. A Large-Scale Benchmark for Vehicle Logo Recognition. In: 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC). IEEE; 2019. p. 479–483.
10. Mizuno S, Okada M, Toriwaki Ji. Virtual sculpting and virtual woodcut printing. The Visual Computer. 1998;2(14):39–51.
11. Mizunoy S, Okadayy M, Toriwakiy J. An interactive designing system with virtual sculpting and virtual woodcut printing. In: Computer Graphics Forum. vol. 18. Wiley Online Library; 1999. p. 183–194.
12. Mizuno S, Kasaura T, Okouchi T, Yamamoto S, Okada M, Toriwaki J. Automatic generation of virtual woodblocks and multicolor woodblock printing. In: Computer Graphics Forum. vol. 19. Wiley Online Library; 2000. p. 51–58.
13. Mizuno S, Okada M, Yamamoto S, Toriwaki Ji. Japanese Traditional Printing" Ukiyo-e" in a Virtual Space. FORMA-TOKYO-. 2001;16(3):233–239.
14. Mizuno S, Okada M, Toriwaki Ji, Yamamoto S. Improvement of the virtual printing scheme for synthesizing Ukiyo-e. In: 2002 International Conference on Pattern Recognition. vol. 3. IEEE; 2002. p. 1043–1046.
15. Mello V, Jung CR, Walter M. Virtual woodcuts from images. In: Proceedings of the 5th international conference on Computer graphics and interactive techniques in Australia and Southeast Asia; 2007. p. 103–109.
16. Li J, Xu D. A Scores Based Rendering for Yunnan Out-of-Print Woodcut. In: 2015 14th International Conference on Computer-Aided Design and Computer Graphics (CAD/Graphics); 2015. p. 214–215.
17. Li J, Xu D. Image stylization for Yunnan out-of-print woodcut through virtual carving and printing. In: International Conference on Technologies for E-Learning and Digital Entertainment. Springer; 2016. p. 212–223.
18. Mesquita DP, Walter M. Synthesis and Validation of Virtual Woodcuts Generated with Reaction-Diffusion. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics. Springer; 2019. p. 3–29.
19. Hertzmann A, Jacobs CE, Oliver N, Curless B, Salesin DH. Image analogies. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2; 2023. p. 557–570.
20. Efros AA, Freeman WT. Image quilting for texture synthesis and transfer. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2; 2023. p. 571–576.
21. Zhu M, Yang M, Meng W, Li P. Sand painting conversion based on detail preservation. Computers & Graphics. 2023;115:371–381.
22. Akita K, Morimoto Y, Tsuruno R. Hand-drawn anime line drawing colorization of faces with texture details. Computer Animation and Virtual Worlds. 2023;p. e2198.
23. Lin H, Xu C, Liu C. FAEC-GAN: An unsupervised face-to-anime translation based on edge enhancement and coordinate attention. Computer Animation and Virtual Worlds. 2023;p. e2135.
24. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27.
25. Mordvintsev A, Olah C, Tyka M. Inceptionism: Going Deeper into Neural Networks 2015. Available from: https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.
26. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2414–2423.
27. Elgammal A, Liu B, Elhoseiny M, Mazzone M. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. arXiv preprint arXiv:170607068. 2017;.
28. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
29. Shen F, Xie Y, Zhu J, Zhu X, Zeng H. Git: Graph interactive transformer for vehicle re-identification. IEEE Transactions on Image Processing. 2023;.
30. Xie Y, Shen F, Zhu J, Zeng H. Viewpoint robust knowledge distillation for accelerating vehicle re-identification. EURASIP Journal on Advances in Signal Processing. 2021;2021:1–13.
31. Xu R, Shen F, Wu H, Zhu J, Zeng H. Dual modal meta metric learning for attribute-image person re-identification. In: 2021 IEEE International Conference on Networking, Sensing and Control (ICNSC). vol. 1. IEEE; 2021. p. 1–6.
32. Hu J, Huang Z, Shen F, He D, Xian Q. A Rubust Method for Roof Extraction and Height Estimation. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. IEEE; 2023. p. 770–771.
33. Weng W, Lin W, Lin F, Ren J, Shen F. A novel cross frequency-domain interaction learning for aerial oriented object detection. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Springer; 2023. p. 292–305.
34. Qiao C, Shen F, Wang X, Wang R, Cao F, Zhao S, et al. A Novel Multi-Frequency Coordinated Module for SAR Ship Detection. In: 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE; 2022. p. 804–811.
35. Shen F, Zhu J, Zhu X, Huang J, Zeng H, Lei Z, et al. An Efficient Multiresolution Network for Vehicle Reidentification. IEEE Internet of Things Journal. 2021;9(11):9049–9059.
36. Wu H, Shen F, Zhu J, Zeng H, Zhu X, Lei Z. A sample-proxy dual triplet loss function for object re-identification. IET Image Processing. 2022;16(14):3781–3789.
37. Shen F, Zhu J, Zhu X, Xie Y, Huang J. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. IEEE Transactions on Intelligent Transportation Systems. 2021;23(7):8793–8804.
38. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:211210741. 2021;.

39. Ding M, Yang Z, Hong W, Zheng W, Zhou C, Yin D, et al. Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems. 2021;34:19822–19835.
40. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems. 2022;35:36479–36494.
41. Gafni O, Polyak A, Ashual O, Sheynin S, Parikh D, Taigman Y. Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. Springer; 2022. p. 89–106.
42. Balaji Y, Nah S, Huang X, Vahdat A, Song J, Kreis K, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:221101324. 2022;.
43. Xue Z, Song G, Guo Q, Liu B, Zong Z, Liu Y, et al. Raphael: Text-to-image generation via large mixture of diffusion paths. arXiv preprint arXiv:230518295. 2023;.
44. Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. Advances in neural information processing systems. 2021;34:8780–8794.
45. Ranftl R, Lasinger K, Hafner D, Schindler K, Koltun V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence. 2020;44(3):1623–1637.
46. Canny J. A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1986;PAMI-8(6):679–698.
47. Zhao S, Chen D, Chen YC, Bao J, Hao S, Yuan L, et al. Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. arXiv preprint arXiv:230516322. 2023;.
48. Shen F, Ye H, Zhang J, Wang C, Han X, Yang W. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. arXiv preprint arXiv:231006313. 2023;.
49. Seguin B, Striolo C, diLenardo I, Kaplan F. Visual link retrieval in a database of paintings. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14. Springer; 2016. p. 753–767.
50. Mao H, Cheung M, She J. Deepart: Learning joint representations of visual arts. In: Proceedings of the 25th ACM international conference on Multimedia; 2017. p. 1183–1191.
51. Achlioptas P, Ovsjanikov M, Haydarov K, Elhoseiny M, Guibas LJ. Artemis: Affective language for visual art. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 11569–11579.
52. Li J, Li D, Xiong C, Hoi S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. PMLR; 2022. p. 12888–12900.
53. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems. 2020;33:6840–6851.
54. Ho J, Salimans T. Classifier-free diffusion guidance. arXiv preprint arXiv:220712598. 2022;.
55. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer; 2015. p. 234–241.
56. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. PMLR; 2021. p. 8748–8763.
57. Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv preprint arXiv:160706450. 2016;.
58. Zhu H, Ke W, Li D, Liu J, Tian L, Shan Y. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 4692–4702.
59. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 22500–22510.
60. Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv preprint arXiv:201002502. 2020;.
61. Ilharco G, Wortsman M, Wightman R, Gordon C, Carlini N, Taori R, et al.. Openclip. OpenAI 2021. https://github.com/mlfoundations/open_clip.
62. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:171105101. 2017;.
63. Hessel J, Holtzman A, Forbes M, Bras RL, Choi Y. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:210408718. 2021;.
64. Xu J, Liu X, Wu Y, Tong Y, Li Q, Ding M, et al. Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems. 2024;36.
65. Yu J, Xu Y, Koh JY, Luong T, Baid G, Wang Z, et al. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:220610789. 2022;2(3):5.

AUTHOR BIOGRAPHY

Yangchunxue Ou received her bachelor's degree in Communication Engineering from Chongqing University of Posts and Telecommunications in 2019 and is studying as a postgraduate at Sichuan Fine Arts Institute. Her research interests include animation, digital media, computer graphics, and deep learning.

Jingjun Xu obtained her master's degree in Design from Sichuan Fine Arts Institute in 2007 and is pursuing a Ph.D. at Chongqing Normal University. She serves as the vice president at the School of Fine Arts Education of Sichuan Fine Arts Institute, focusing primarily on research areas such as film and television animation, virtual simulation, and interaction.