DiffQRCoder: Diffusion-based Aesthetic QR Code Generation with Scanning Robustness Guided Iterative Refinement

Jia-Wei Liao[1,2] Winston Wang[1*] Tzu-Sian Wang[1*] Li-Xuan Peng[1*] Ju-Hsuan Weng[1,2]
Cheng-Fu Chou[2] Jun-Cheng Chen[1]

[1] Research Center for Information Technology Innovation, Academia Sinica,
[2] National Taiwan University

Original QR Code

Winter wonderland, fresh snowfall, evergreen trees, cozy log cabin, smoke rising from chimney, aurora borealis in night sky.

Cherry blossom festival, pink petals floating in the air, traditional lanterns, peaceful river, people in kimonos, sunny day.

Majestic waterfall, lush rainforest, rainbow in the mist, exotic birds, vibrant flowers, serene pool below.

Abandoned amusement park, overgrown rides, haunting beauty, sense of nostalgia, sunset lighting.

Figure 1. **Aesthetic QR codes generated from DiffQRCoder**. Our method takes a QR code and a text prompt as input to generate an aesthetic QR code. We leverage the pre-trained ControlNet and guide the generation process using our proposed Scanning Robust Perceptual Guidance (SRPG) to ensure the generated code is both scannable and attractive.

Abstract

With the success of Diffusion Models for image generation, the technologies also have revolutionized the aesthetic Quick Response (QR) code generation. Despite significant improvements in visual attractiveness for the beautified codes, their scannabilities are usually sacrificed and thus hinder their practical uses in real-world scenarios. To address this issue, we propose a novel training-free ***Diff****usion-based* ***QR Code*** *generato****r*** *(DiffQRCoder) to effectively craft both scannable and visually pleasing QR codes. The proposed approach introduces Scanning-Robust Perceptual Guidance (SRPG), a new diffusion guidance for Diffusion Models to guarantee the generated aesthetic codes to obey the ground-truth QR codes while maintaining their attractiveness during the denoising process. Additionally, we present another post-processing technique, Scanning Robust Manifold Projected Gradient Descent (SR-MPGD), to further enhance their scanning robustness through iterative latent space optimization. With extensive experiments, the results demonstrate that our approach not only outperforms other compared methods in Scanning Success Rate (SSR) with better or comparable CLIP aesthetic score (CLIP-aes.) but also significantly improves the SSR of the ControlNet-only approach from 60% to 99%. The subjective evaluation indicates that our approach achieves promising visual attractiveness to users as well. Finally, even with different scanning angles and the most rigorous error tolerance settings, our approach robustly achieves over 95% SSR, demonstrating its capability for real-world applications. Our project page is available at* https://jwliao1209.github.io/DiffQRCoder.

1. Introduction

Quick Response (QR) [7] codes are ubiquitous in daily transactions, information sharing, and marketing, driven by their quick readability and the widespread use of smartphones. However, the standard black-and-white QR codes lack visual appeal. Aesthetic QR codes offer a solution by not only capturing user attention but also seamlessly integrating with product designs, enhancing user experiences,

* denotes equal contribution.

2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3315-1083-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/WACV61041.2025.00577

2642-9381/25/$31.00 ©2025 IEEE
DOI 10.1109/WACV61041.2025.00577

Figure 2. **Existing methods struggle to balance scannability and aesthetics.** Although QRBTF [18] generate visually appealing QR codes, they lack scanning robustness. Conversely, QR Code AI Art [21] and QR Diffusion [15] produce better scanning robust QR codes but are visually less appealing. Our approach can generate both attractive and scannable QR codes. Red frames indicate unscannable codes, while green frames denote scannable codes. Zoom in for better viewing details.

and amplifying marketing effectiveness. By creating visually appealing QR codes, businesses can elevate brand engagement and improve advertising impact, making them a valuable tool for both functionality and design. Recognizing the commercial value of aesthetic QR codes, numerous beautification techniques have been thus developed.

For this purpose, some previous works have attempted to generate aesthetic QR codes via style-transfer-based techniques [39] to blend style textures with the QR code patterns. However, these methods often lack flexibility and can reduce scanning robustness. Instead, current prevailing commercial products [24] have adopted generative models to create stylized QR codes, primarily employing Diffusion Models (e.g., ControlNet [45]). The mainstream methodology is to adjust the Classifier-Free Guidance (CFG) weights [6] in ControlNet to create visually pleasing QR codes. However, selecting CFG weights presents a trade-off between scannability and visual quality (Fig. 2). In practical applications, manual post-processing is often used to fix unscannable codes, but this process is time-consuming and labor-intensive. Therefore, it is still an open challenge to generate aesthetic QR codes with a good balance between visual attractiveness and scanning robustness.

To address the instability of scanability in previous generative-based methods, we propose **Diff**usion-based **QR Code** generato**r** (DiffQRCoder), a training-free approach to balance the scannability and aesthetics of QR codes. We introduce Scanning Robust Loss (SRL), specifically designed for evaluating the scannability of a beautified QR code with respect to its reference code. Building on SRL, we develop Scanning Robust Perceptual Guidance (SRPG), an extension of the Classifier Guidance concept [2, 6, 44], which ensures generation fidelity to ground-truth QR codes while preserving aesthetics during the denoising process. Besides, we develop a post-processing technique called Scanning Robust Manifold Projected Gradient Descent (SR-MPGD) further to enhance the scanning robustness through iterative latent space optimization utilizing a pre-trained Variational Autoencoder (VAE) [17]. Specifically, our framework features the following key designs: 1) The proposed framework is training-free and compatible with existing diffusion models. 2) Our approach exploits the error tolerance capability inherent in standard QR codes for more flexible and precise manipulation over QR code image pixels.

Finally, extensive experiments demonstrate that our approach outperforms other compared models in Scanning Success Rate (SSR) with better or comparable CLIP aesthetic scores (CLIP-aes.) [32]. Specifically, it achieves 99% SSR while better preserving CLIP aesthetic scores.

Our main contributions are summarized as follows:

1. We propose a two-stage iterative refinement framework with a novel Scanning Robust Perceptual Guidance (SRPG) tailored for QR code mechanisms to generate scanning-robust and visually appealing aesthetic QR codes without training.

2. We propose Scanning Robust Manifold Projected Gradient Descent (SR-MPGD) for post-processing, enabling Scanning Success Rate (SSR) of aesthetic QR code up to 100% through latent space optimization.

3. Extensive quantitative and qualitative experiments demonstrate that our proposed framework significantly enhances the Scanning Success Rate (SSR) of the ControlNet-only approach from 60% to nearly 100%, without compromising aesthetics. User subjective evaluations further confirm the visual appeal of our QR codes.

2. Related Works

2.1. Image Diffusion Models

Recently, Diffusion Models [12, 33] have emerged as powerful generative models, demonstrating superior unconditional image generation capabilities compared to GAN-based models [6, 9]. Dhariwal et al. [6] introduced the concept of Classifier Guidance to control the sampling process via gradients from pre-trained classifiers, which has since been further developed with more generalized conditional probability terms for more freely control [1, 16, 20, 44, 48].

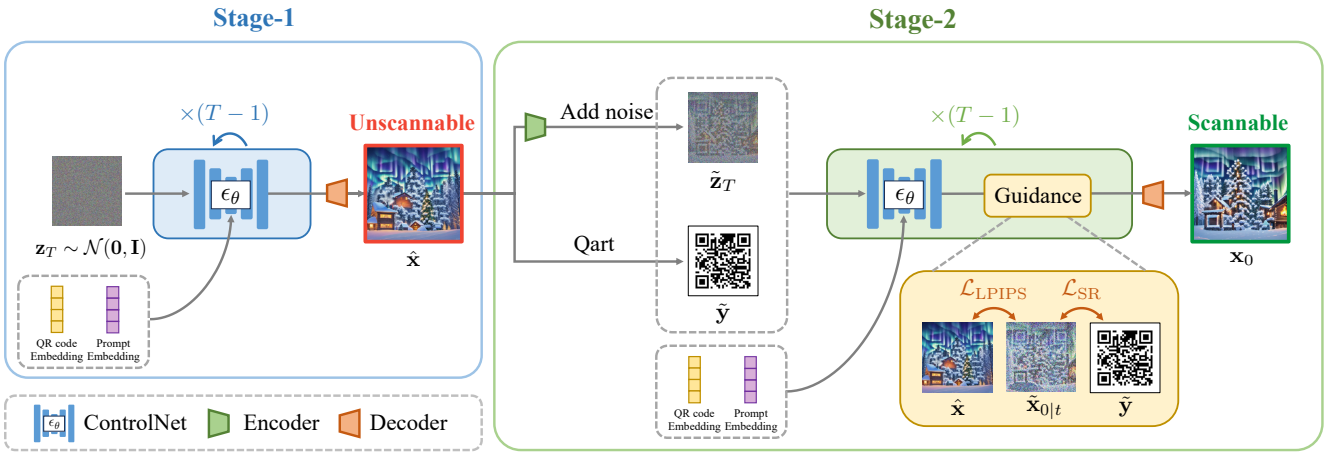

Figure 3. **An overview of our two-stage pipeline with Scanning Robust Perceptual Guidance (SRPG).** First, we encode target QR code $\mathbf{y}$ and prompt p to embeddings for ControlNet input. In Stage-1, we utilize the pre-trained ControlNet to generate an attractive yet unscannable QR code. In Stage-2, we convert the QR code from Stage-1 into a latent representation $\tilde{z}_T$ by adding Gaussian noise and transforming the $\mathbf{y}$ to $\tilde{\mathbf{y}}$, which has a more similar pattern as $\hat{\mathbf{x}}$, using Qart [5]. Finally, we feed the latent and the transformed code into ControlNet, guided by Scanning Robust Perceptual Guidance (SRPG), to create an aesthetic QR code with scannability.

However, Diffusion Models require substantial computational resources, especially for high-resolution images. To address this issue, Rombach et al. [30] proposed the Latent Diffusion Model (LDM), leveraging a pre-trained VAE to compress high-resolution images into a lower-dimensional latent space. This approach enhances efficiency in the diffusion process while preserving visual quality. For more fine-grained manipulations in downstream tasks, Zhang et al. [46], Qin et al. [27], and Zavadski et al. [45] proposed adaptation strategies that allow users to fine-tune only the extra output layer instead of the entire model.

These advancements have significantly impacted fields including image editing [4, 6, 10, 11, 22, 23, 25, 43], text-to-image synthesis [29–31], and commercial product development, exemplified by DALL-E2 [26] and Midjourney [14].

2.2. Aesthetic QR Codes

2.2.1 Non-generative-based Models

Previous works on aesthetic QR codes have focused on three main techniques: module deformation, module reshuffling, and style transfers. Module-deformation methods, such as Halftone QR codes [3], integrate reference images by deforming and scaling code modules while maintaining scanning robustness. Module-reshuffling, introduced by Qart [5], rearranges code modules using Gaussian-Jordan elimination to align pixel distributions with reference images to ensure decoding accuracy. Image processing techniques have also been developed to enhance visual quality, such as region of interest [41], central saliency [19], and global gray values [42]. Xu et al. [42] proposed Stylized aEsthEtic (SEE) QR codes, pioneering style-transfer-based techniques for aesthetic QR codes but encountered visual artifacts from pixel clustering. ArtCoder [39] reduced these artifacts by optimizing style, content, and code losses jointly, although some artifacts remain. Su et al. [38] further improved aesthetics with the Module-based Deformable Convolutional Mechanism (MDCM). However, these techniques require reference images, which leads to a lack of flexibility and variation.

2.2.2 Generative-based Models

With the rise of diffusion-based image manipulation and conditional control techniques, previous works such as QR Diffusion [15], QR Code AI Art [21] and QRBTF [18] have leveraged generative power of diffusion models to create aesthetic QR codes, primarily relying on ControlNet [45] for guidance. However, more fine-grained guidance that aligns with the inherent mechanisms of QR codes remains unexplored. Another non-open-source method Text2QR [40], introduced a three-stage pipeline that first generates an unscannable QR code using pre-trained ControlNet, followed by their proposed SELR independent of the diffusion sampling process to ensure scannability. However, the diffusion sampling process of Text2QR can not guarantee scannability on its own; our approach aims to fill this gap by designing a training-free, fine-grained guidance integrating scannability criteria into the diffusion sampling process.

3. Method

DiffQRCoder is designed to generate a scannable and attractive QR code from a given text prompt p and a target QR code $\mathbf{y}$, which consists of $m \times m$ modules, each of size $s \times s$ pixels. Fig. 3 illustrates the overall architecture of

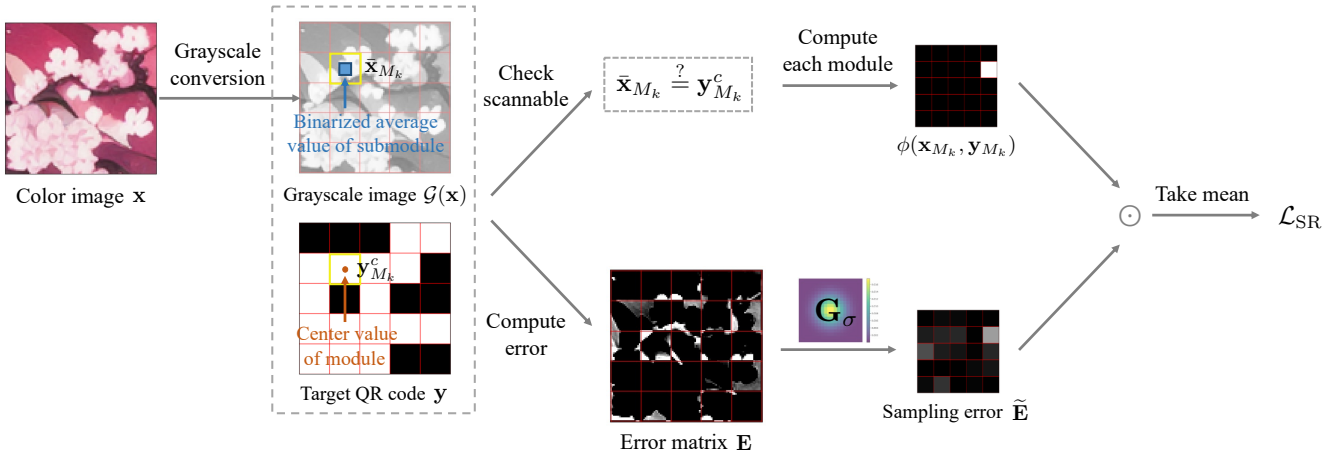

Figure 4. **An illustration of Scanning Robust Loss (SRL).** SRL is designed at the module level, tailored to the QR code mechanism. The process begins by constructing a pixel-wise error matrix that measures the differences between the pixel values of the target QR code and the grayscale image. Subsequently, the error for each module is re-weighted using a Gaussian kernel, and the central submodule is extracted to implement an early-stopping mechanism. The mechanism stops refining the module and evaluating its error once the average pixel value of the central submodule matches the center pixel value of the target module. Finally, SRL can be calculated as the average error across all modules in the code.

DiffQRCoder, consisting of two stages. In Stage-1, ControlNet [45] without our proposed guidance is employed to create a visually pleasing yet unscannable QR code. In Stage-2, we convert the QR code into a latent representation by adding Gaussian noise and transform the $\mathbf{y}$ to $\tilde{\mathbf{y}}$, which has a more similar pattern as $\hat{\mathbf{x}}$, using Qart [5]. We then feed the latent and the transformed code into ControlNet with Scanning Robust Perceptual Guidance (SRPG). The guided loss function includes perceptual loss and our proposed Scanning Robust Loss (SRL), ensuring that the generated QR codes are both scannable and attractive. Besides, we propose a post-processing technique called Scanning Robust Manifold Perceptual Gradient Descent (SR-MPGD) to boost the scanning robustness. Detailed descriptions on SRL, the two-stage pipeline, and SR-MPGD are provided in Sec. 3.1, Sec. 3.2, and Sec. 3.3, respectively.

3.1. Scanning Robust Loss

SRL (Fig. 4) is designed to assess the scannability of a beautified QR code with respect to its target code, and aims to provide the guidance signal for image manipulation at the module level. It begins with an error matrix that evaluates the differences between the pixel values of the target code and the image in grayscale. Next, the matrix is re-weighted to account for varying scanning probabilities across the code. Consequently, we extract the central submodule within each module due to its importance in decoding. Additionally, an early-stopping is implemented to prevent over-optimization.

Pixel-wise Error Matrix. Given a normalized image $\mathbf{x}$, a target QR code $\mathbf{y}$, and a grayscale conversion operator $\mathcal{G}$ (cf., Appendix A). We formulate a pixel-wise error matrix $\mathbf{E}$ that calculates the differences in pixel values between $\mathbf{y}$ and $\mathcal{G}(\mathbf{x})$. $\mathbf{E}$ is calculated as follows:

$$\begin{aligned}\mathbf{E} = &\max(1 - 2\mathcal{G}(\mathbf{x}), 0) \odot \mathbf{y} + \\ &\max(2\mathcal{G}(\mathbf{x}) - 1, 0) \odot (1 - \mathbf{y}),\end{aligned} \tag{1}$$

where $\max(\cdot, \cdot)$ operator is applied pixel-wisely, and $\odot$ denotes the Hadamard product. The first term in the equation addresses the white squares of the QR code, while the second term focuses on the black squares.

Error Re-weighting. Not all pixels are equally likely to be scanned. According to ART-UP [41], the scanning probabilities of pixels within each module follow a Gaussian distribution. This implies that pixels closer to the center of each module are more important. Consequently, we re-weight the module error M_k as

$$\widetilde{E}_{M_k} = \sum_{(i,j) \in M_k} \mathbf{G}_\sigma(i,j) \cdot \mathbf{E}(i,j), \tag{2}$$

where (i, j) indicates the coordinate of a pixel in M_k, and $\boldsymbol{G}_\sigma$ is a Gaussian kernel function with standard deviation $\sigma = \lfloor \frac{s-1}{5} \rceil$.

Central Submodule Filter. ZXing [37], a popular barcode scanning library, notes that only the central pixels of

each module are essential for decoding QR codes. According to Chu et al. [3], each module is divided into 3×3 submodules. They observed that a QR code remains scannable if the binarized average pixel value in the central submodule matches the center pixel value of the target module .This observation enables the creation of visual variations in the peripheral submodules.

A central submodule filter $\mathbf{F}$ is applied to extract the central submodule:

$$\mathbf{F} = \frac{1}{\lceil \frac{m}{3} \rceil^2} \begin{bmatrix} \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{\lceil \frac{m}{3} \rceil \times \lceil \frac{m}{3} \rceil} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} \end{bmatrix}_{m \times m}. \quad (3)$$

The binarized average pixel value of the center submodule $\mathbf{x}_{M_k}$ is calculated as:

$$\bar{\mathbf{x}}_{M_k} = \mathbb{I}_{[\frac{1}{2},1]} \left(\sum_{(i,j) \in M_k} \mathbf{F}(i,j) \cdot \mathcal{G}(\mathbf{x}_{M_k}(i,j)) \right), \quad (4)$$

where $\mathbb{I}_A$ is an indicator function of set A.

To determine whether $\mathbf{x}_{M_k}$ is correctly matched, we define the function ϕ as:

$$\phi(\mathbf{x}_{M_k}, \mathbf{y}_{M_k}) = \begin{cases} 0, & \bar{\mathbf{x}}_{M_k} = \mathbf{y}^c_{M_k}, \\ 1, & \bar{\mathbf{x}}_{M_k} \neq \mathbf{y}^c_{M_k}, \end{cases}, \quad (5)$$

where $\mathbf{y}^c_{M_k}$ represents the center pixel value of the target module.

Early-stopping Mechanism. We employ an early-stopping mechanism at the module level to prevent over-optimization. This mechanism stops refining a module once it can be correctly decoded, i.e. $\phi(\mathbf{x}_{M_k}, \mathbf{y}_{M_k}) = 0$. ϕ acts as a switch that determines whether to update $\mathbf{x}_{M_k}$, hence its gradient will not be used to update $\mathbf{x}_{M_k}$. Therefore, we use the stop gradient operator $\mathrm{sg}[\cdot]$ to detach this term from the computation graph. The SRL can be expressed as:

$$\mathcal{L}_{\text{SR}}(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{k=1}^{N} \phi(\mathrm{sg}[\mathbf{x}_{M_k}], \mathbf{y}_{M_k}) \cdot \widetilde{E}_{M_k}, \quad (6)$$

where N is the number of modules.

3.2. Two-stage Pipeline with Scanning Robust Perceptual Guidance

DiffQRCoder utilizes a two-stage pipeline. In Stage-1, we use ControlNet, without our proposed guidance, to create a visually appealing yet unscannable QR code. In Stage-2, we refine the generation process with Scanning Robust Perceptual Guidance (SRPG). This guidance employs a loss function that combines Learned Perceptual Image Path Similarity (LPIPS) [47](cf. Appendix B.1), denote as $\mathcal{L}_{\text{LPIPS}}$, and $\mathcal{L}_{\text{SR}}$, ensuring a balance between aesthetics and scanning robustness.

3.2.1 Stage-1

In stage-1, we first encode p as the prompt embedding $\mathbf{e}_p$ and $\mathbf{y}$ as the QR code embedding $\mathbf{e}_{\text{code}}$. And we sample a noise latent $\mathbf{z}_t$ from a standard normal distribution. Then feed into ControlNet to generate an unscannable QR code $\hat{\mathbf{x}}$, which will be used in stage-2 for perceptual regularizing reference.

3.2.2 Stage-2

In stage-2 we adopt the unscannable QR code $\hat{\mathbf{x}}$ generated in stage-1 as a starting point for enhancing scannability and a regularizing reference for LPIPS to preserve aesthetics. First, we convert image $\hat{\mathbf{x}}$ into a latent representation $\tilde{\mathbf{z}}_t$ using the VAE encoder and by adding Gaussian noise. We also transform the target QR code $\mathbf{y}$ to be a more similar pattern as $\hat{\mathbf{x}}$ for better conditioning. Both $\tilde{\mathbf{z}}_t$ and $\tilde{\mathbf{y}}$ are then fed into ControlNet. The predicted clean latent at each timestep t can be calculated as

$$\tilde{\mathbf{z}}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\tilde{\mathbf{z}}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(\tilde{\mathbf{z}}_t, t, \mathbf{e}_p, \mathbf{e}_{\text{code}}) \right), \quad (7)$$

where ϵ_θ denotes the noise predictor of ControlNet.

Since $\mathcal{L}_{\text{SR}}$ and $\mathcal{L}_{\text{LPIPS}}$ operate in the pixel space, we use the pre-trained image decoder of ControlNet, $\mathcal{D}_\theta(\cdot)$, to map $\tilde{\mathbf{z}}_{0|t}$ into the pixel space: $\tilde{\mathbf{x}}_{0|t} = \mathcal{D}_\theta\left(\tilde{\mathbf{z}}_{0|t}\right)$. As a result, the guidance function F_{SRP} can be formulated as:

$$F_{\text{SRP}}(\tilde{\mathbf{z}}_t, \tilde{\mathbf{y}}, \hat{\mathbf{x}}) = \lambda_1 \mathcal{L}_{\text{SR}}(\tilde{\mathbf{x}}_{0|t}, \tilde{\mathbf{y}}) + \lambda_2 \mathcal{L}_{\text{LPIPS}}(\tilde{\mathbf{x}}_{0|t}, \hat{\mathbf{x}}), \quad (8)$$

where λ_1 and λ_2 denote the guidance scales.

Song et al. [35, 36] established a connection between the score function and the estimated noise function, demonstrating that

$$\epsilon_\theta(\mathbf{z}_t, t, \mathbf{e}_p, \mathbf{e}_{\text{code}}) = -\sqrt{1 - \bar{\alpha}_t} \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t). \quad (9)$$

Inspired by [6], the guided noise prediction becomes:

$$\hat{\epsilon}_t = \epsilon_\theta(\tilde{\mathbf{z}}_t, t, \mathbf{e}_p, \mathbf{e}_{\text{code}}) + \sqrt{1 - \bar{\alpha}_t} \nabla_{\tilde{\mathbf{z}}_t} F_{\text{SRP}}(\tilde{\mathbf{z}}_t, \mathbf{y}). \quad (10)$$

Finally, we employ the DDIM sampling [34]:

$$\tilde{\mathbf{z}}_{t-1} = \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}} \left(\tilde{\mathbf{z}}_t - \sqrt{1 - \bar{\alpha}_t} \hat{\epsilon}_t \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \hat{\epsilon}_t. \quad (11)$$

After T iterations, we decode $\tilde{\mathbf{z}}_0$ into the pixel space by $\mathcal{D}_\theta(\cdot)$ to obtain the generated aesthetic QR code $\mathbf{x}_0$. The complete algorithm for our two-stage generation pipeline is provided in Appendix C.2. Detailed derivations of the formulas can be found in Appendix B.2 and B.3.

3.3. Post-processing with Scanning-Robust Manifold Projected Gradient Descent (SR-MPGD)

SR-MPGD is a post-processing technique proposed to enhance scanning robustness further. Our goal is to minimize $\mathcal{L}_{\text{SR}}(\mathbf{x}, \mathbf{y})$ while ensuring the refined QR code $\mathbf{x}$ still lies on nature image manifold $\mathcal{M}$. The optimization problem is defined as: $\min_{\mathbf{x}\in\mathcal{M}} \mathcal{L}_{\text{SR}}(\mathbf{x}, \mathbf{y})$. This constrained optimization problem can be solved via the Projected Gradient Descent (PGD) algorithm. Inspired by the manifold-preserving nature proposed in [10], we use the pre-trained VAE encoder $\mathcal{E}(\cdot)$ to project the image to its space and then iteratively refine the latent by:

$$\mathbf{z}_0^i = \mathbf{z}_0^{i-1} - \gamma \nabla_{\mathbf{z}} \mathcal{L}_{\text{SR}}(\mathcal{D}(\mathbf{z}_0^{i-1}), \mathbf{y}). \tag{12}$$

Note that $\mathbf{z}_0^i$ indicates the clean image latent output from Sec. 3.2 and in each iterative refinement step, the VAE decoder $\mathcal{D}(\cdot)$ will project the latent back to image manifold. However, the VAE is imperfect and may introduce reconstruction errors in practice. To mitigate this, we incorporate LPIPS loss as a regularization term with a weight $\lambda > 0$:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \mathbf{x}_0) = \mathcal{L}_{\text{SR}}(\mathbf{x}, \mathbf{y}) + \lambda \mathcal{L}_{\text{LPIPS}}(\mathbf{x}, \mathbf{x}_0). \tag{13}$$

The rationale for employing LPIPS loss is to facilitate $\mathcal{L}_{\text{SR}}$ to refine incorrect modules while preserving coarse-grained semantics. Finally, the update rule becomes:

$$\mathbf{z}_0^i = \mathbf{z}_0^{i-1} - \gamma \nabla_{\mathbf{z}_0} \mathcal{L}(\mathcal{D}(\mathbf{z}_0^{i-1}), \mathbf{y}, \mathbf{x}_0). \tag{14}$$

By incorporating this latent optimization, it can converge to local minima against $\mathcal{L}_{\text{SR}}$ near initial latent $\mathbf{z}_0$.

4. Experiments

4.1. Experimental Settings

Implementation Details. In our experiments, 100 text prompts are generated by GPT-4 as the conditions for Stable Diffusion, consistently using `easynegative` as the negative prompt. We employ `Cetus-Mix Whalefall` as the checkpoint for Stable Diffusion [30] and QR code Monster v2 [24] as the checkpoint for ControlNet [46], with the `guidance_scale` set to 1.35 for the latter. We compare with QR code AI Art [21], QR Diffusion [15], and QRBTF [18]. The reason we don't compare with non-generative-based methods is that they rely on reference aesthetic images as input, here we only have prompts describing the aesthetic scenes. And other previous works are unavailable for fair comparison due to non-open-source. Compared models are accessed through their web API using their recommended settings. Detailed parameter settings are provided in Appendix D.

In our QR code setups, we use Version 3 QR codes configured with a medium (M) error correction level and mask pattern 4. Each code includes an 80-pixel padding, and each module is in the size of 20×20 pixels. Additionally, the text message `Thanks reviewer!` is encoded into our QR codes for most experiments in the paper. We also conduct quantitative and qualitative results in different messages for QR code generation. We conduct our experiments using a single NVIDIA RTX 4090 GPU. Generating aesthetic QR codes takes approximately 14 to 18 seconds in our two-stage pipeline, each with 40 inference steps.

Evaluation Metrics. We use qr-verify [8] to measure Scanning Success Rate (SSR) of aesthetic QR codes. For the quantitative assessment of aesthetics, we use CLIP aesthetic score predictor [32] to reflect image quality and visual appeal. This score is referred to as the CLIP aesthetic score (CLIP-aes). We also adopt CLIP-score [28] to assess the text-image alignment of generated aesthetic QR codes.

4.2. Comparison with Other Generative-based Methods

Method	SSR ↑	CLIP-aes. ↑	CLIP-score ↑
QR Code AI Art [13]	90%	5.7003	0.2341
QR Diffusion [15]	96%	5.5150	0.2780
QRBTF [18]	56%	**7.0156**	**0.3033**
DiffQRCoder (Ours)	**99%**	6.8233	0.2992

Table 1. Quantitative comparison with other generative-based methods. DiffQRCoder significantly outperforms other methods in SSR with only an insignificant decrease in CLIP-aes.

Quantitative Results. As shown in Tab. 1, we present the quantitative results of our method compared to previous generative-based methods. Our method outperforms QR Diffusion [15] and QR Code AI Art [13] in SSR, CLIP aesthetic score, and CLIP-score. Compared with QRBTF [18], our method significantly enhances the SSR, albeit with a little trade-off with CLIP aesthetic score [32]. Notably, the text-image alignment measured by CLIP-score shows that our method is close to QRBTF, indicating our method adheres to prompts without distortion.

Furthermore, we test the robustness of our QR codes under various scenarios, including different simulated scanning angles, error correction level configurations, and scanners from multiple devices and open-source software. As presented in Tab. 2, our approach achieves a 97% SSR even at a 45° tilt (Implementation details are provided in Appendix D.1); As presented in Tab. 3 for different error correction levels, our approach still achieves a 96% SSR under the most rigorous setting (7% tolerance), we also provide qualitative results in Appendix; As presented in Tab. 4, for scanning with different scanners, results show that our method can still achieving over 88% SSR even in the worst case. Additionally, we generate QR codes with various encoded messages and assess their SSR, the results are pro-

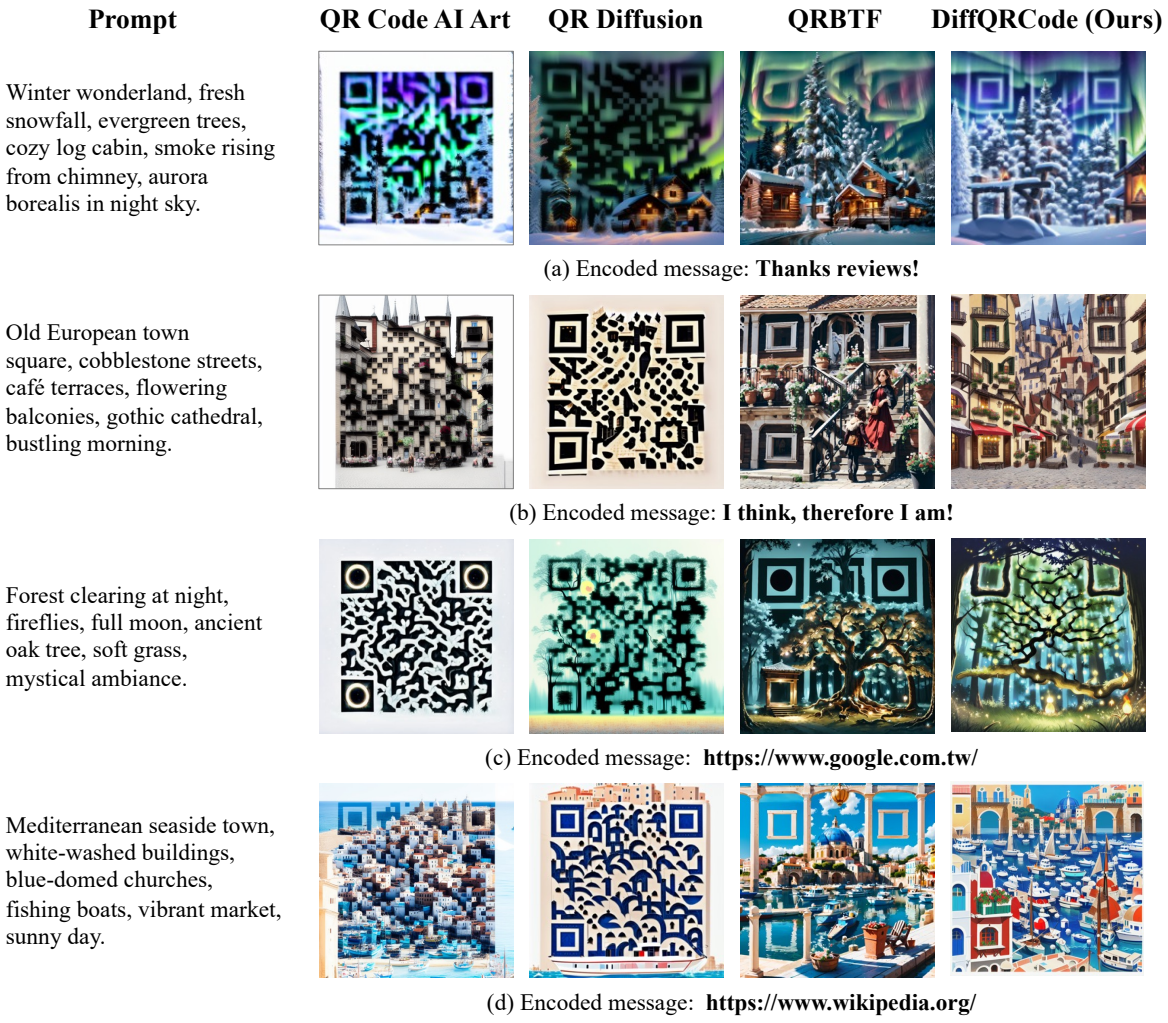

Figure 5. **Qualitative comparison with other generative-based methods.** DiffQRCoder can generate attractive and scannable QR codes with different encoded messages and prompts.

vided in Tab. 5, and their qualitative results are provided in the Appendix.

Degree	0°	15°	30°	45°
SSR ↑	100%	100%	100%	97%

Table 2. Scannability of different angles.

Level	L (7%)	M (15%)	Q (25%)	H (30%)
SSR ↑	96%	100%	100%	100%

Table 3. Scannability of different QR code error correction level.

Device	**qr-verify**	**iPhone 13**	**Pixel 7**
SSR ↑	100.00%	97%	88%

Table 4. Scannability results of different devices.

Qualitative Results. We present qualitative comparisons with previous methods in Fig. 5. Compared to QR Code AI Art and QR Diffusion, our method exhibits a more harmonized pattern blending with concepts in prompts; Compared to QRBTF, our method trades little aesthetics for scannability, achieving more scanning-robust QR code generation than ControlNet-only methods. Fig. 6 illustrates the reduction of error modules during the iterative refinement process using our SRPG method. The error modules, highlighted in red, progressively diminish as the denoising step advances. Once the error level falls below a tolerable threshold, the QR code becomes scannable. Moreover, we present qualitative results for a given QR code in different prompts in Fig. 1. We provide more qualitative results in Appendix.

Message	**SSR** ↑
`I think, therefore I am.`	97%
`You are the apple of my eye.`	100%
`https://www.google.com.tw/`	100%
`https://www.wikipedia.org/`	97%

Table 5. Scannability of different QR code encoded messages.

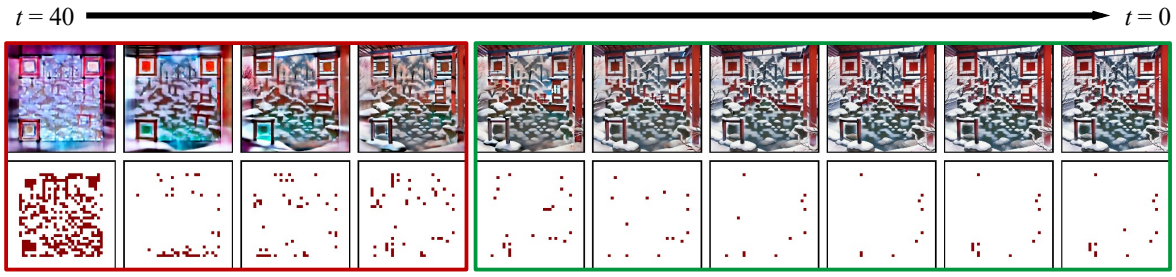

Figure 6. Visualization of $\mathbf{x}_{0|t}$ and its error modules during sampling steps.

Subjective Results. We conducted a user-subjective aesthetic preference study with 387 participants, the result is reported in Tab. 6. Although QRBTF [18] ranked first, our method closely follows with little difference. Considering the limited scannability of QRBTF, which achieved only 56%, our approach is the leading method for effectively balancing visual attractiveness with scannability. The details of the average rank calculation and questionnaire design are provided in the Appendix E.

Methods	Average Rank ↓	SSR ↑
QR Code AI Art [13]	2.71	90%
QR Diffusion [15]	3.18	96%
QRBTF [18]	**1.86**	56%
DiffQRCoder (Ours)	2.25	**99%**

Table 6. The weighted aesthetic ranks for different methods.

4.3. Ablation Studies

Effectiveness of Different SRPG Guidance Scales. In this study, we investigate the effectiveness of our proposed $\mathcal{L}_{\text{SR}}$ and regularizing $\mathcal{L}_{\text{LPIPS}}$ respectively (cf., Eq. 8). In Tab. 7, Stage-1-only indicates only ControlNet is adopted to generate aesthetic QR codes. First, we fix $\lambda_2 = 0$, and only perform Stage-2 generation with SRPG. In the absence of $\hat{\mathbf{x}}$, we use ControlNet with original $\mathbf{y}$ and text prompts to generate images. As shown in upper half of Tab. 7, increasing λ_1 significantly improves SSR while slightly decreasing CLIP aesthetic score. Second, we fix $\lambda_1 = 500$ and perform a full two-stage generation with SRPG. As shown in lower half of Tab. 7, increasing λ_2 improves CLIP aesthetic score while preserving SSR.

Effectiveness of SR-MPGD. SR-MPGD is a post-processing technique designed to enhance scanning robustness further. In our experiments, we set step size $\gamma = 1000$ (cf., Eq. 14), and LPIPS $\lambda = 0.01$(cf., Eq. 13). As reported in Tab. 7, it substantially improves SSR with only a negligible decrease in CLIP aesthetic score. By implementing SR-MPGD, we can even achieve 100% SSR in certain cases.

Stage	λ_1	λ_2	SR-MPGD	CLIP-aes. ↑	SSR ↑
Stage-1-only	-	-		7.0661	60%
Two-stage	400	0		6.7860	86%
Two-stage	500	0		6.7259	88%
Two-stage	600	0		6.7183	94%
Two-stage	1000	0		6.5667	93%
Two-stage	400	0	✓	6.7567	98%
Two-stage	500	0	✓	6.7097	100%
Two-stage	600	0	✓	6.7002	99%
Two-stage	1000	0	✓	6.5629	99%
Two-stage	500	2		6.8600	90%
Two-stage	500	3		6.8744	89%
Two-stage	500	5		6.8357	89%
Two-stage	500	10		6.8409	88%
Two-stage	500	2	✓	6.8204	98%
Two-stage	500	3	✓	6.8233	99%
Two-stage	500	5	✓	6.7779	100%
Two-stage	500	10	✓	6.8040	97%

Table 7. Ablations for our proposed pipeline.

5. Conclusion

In this paper, we introduce a novel training-free Diffusion-based QR Code generator (DiffQRCoder). We propose Scanning-Robust Loss (SRL) to enhance QR code scannability and establish its connection with Scanning-Robust Perceptual Guidance (SRPG). Our two-stage generation pipeline with iterative refinement integrates SRPG to produce aesthetic QR codes. Additionally, we introduce Scanning-Robust Manifold Projected Gradient Descent (SR-MPGD) to further ensure scannability. Compared to existing methods, our approach significantly improves SSR without compromising visual appeal and is competent for real-world applications.

6. Acknowledgements

This research is supported by National Science and Technology Council, Taiwan (R.O.C), under the grant number of NSTC-112-2634-F-002-006, NSTC-112-2222-E-001-001-MY2, NSTC-113-2634-F-001-002-MBK, NSTC-113-2221-E-002-201 and Academia Sinica under the grant number of AS-CDA-110-M09. We sincerely thank Ernie Chu for his inspiring discussions and valuable feedback.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 2

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *ICLR*, 2024. 2

[3] Hung-Kuo Chu, Chia-Sheng Chang, Ruen-Rone Lee, and Niloy J. Mitra. Halftone qr codes. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 2013. 3, 5

[4] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2022. 3

[5] R. Cox. Qartcodes, 2012. 3, 4

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 2, 3, 5

[7] International Organization for Standardization. Information technology automatic identification and data capture techniques code symbology QR Code, 2000. 1

[8] Anthony Fu. qr-verify. https://github.com/antfu/qr-verify, 2023. 6

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 2020. 2

[10] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, et al. Manifold preserving guided diffusion. In *ICLR*, 2023. 3, 6

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2022. 3

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[13] huggingface projects. Qr-code ai art, 2023. 6, 8

[14] Midjourney Inc. Midjourney, 2023. 3

[15] QR Diffusion Inc. Qr diffusion, 2024. 2, 3, 6, 8

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 2

[17] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[18] IoC Lab. Qrbtf, 2023. 2, 3, 6, 8

[19] Shih-Syun Lin, Min-Chun Hu, Chien-Han Lee, and Tong-Yee Lee. Efficient qr code beautification with high quality visual content. *IEEE Transactions on Multimedia*, 2015. 3

[20] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *WACV*, 2023. 2

[21] Thibaut Melen and Nicolas. Qr code ai, 2023. 2, 3, 6

[22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 3

[23] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 3

[24] monster labs. Qr code monster, 2023. 2, 6

[25] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 3

[26] OpenAI. Dall-e-2, 2023. 3

[27] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In *NeurIPS*, 2023. 3

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 6

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 2, 6

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 2

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 5

[35] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019. 5

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 5

[37] Open Source. Zxing: zebra crossing, 2013. 4

[38] Hao Su, Jianwei Niu, Xuefeng Liu, Qingfeng Li, Ji Wan, and Mingliang Xu. Q-art code: Generating scanning-robust art-style qr codes by deformable convolution. In *ACMMM*, 2021. 3

[39] Hao Su, Jianwei Niu, Xuefeng Liu, Qingfeng Li, Ji Wan, Mingliang Xu, and Tao Ren. Artcoder: an end-to-end method for generating scanning-robust stylized qr codes. In *CVPR*, 2021. 2, 3

[40] Guangyang Wu, Xiaohong Liu, Jun Jia, Xuehao Cui, and Guangtao Zhai. Text2qr: Harmonizing aesthetic customization and scanning robustness for text-guided qr code generation. *arXiv preprint arXiv:2403.06452*, 2024. 3

[41] Mingliang Xu, Qingfeng Li, Jianwei Niu, Hao Su, Xiting Liu, Weiwei Xu, Pei Lv, Bing Zhou, and Yi Yang. Art-up: A novel method for generating scanning-robust aesthetic qr codes. *ACM TOMM*, 2021. 3, 4

[42] Mingliang Xu, Hao Su, Yafei Li, Xi Li, Jing Liao, Jianwei Niu, Pei Lv, and Bing Zhou. Stylized aesthetic qr code. *IEEE Transactions on Multimedia*, 2019. 3

[43] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *NeurIPS*, 2024. 3

[44] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *ICCV*, 2023. 2

[45] Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Designing an efficient and effective architecture for controlling text-to-image diffusion models. *arXiv preprint arXiv:2312.06573*, 2023. 2, 3, 4

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3, 6

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[48] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *NeurIPS*, 2022. 2