PTDiffusion: Free Lunch for Generating Optical Illusion Hidden Pictures with Phase-Transferred Diffusion Model

Xiang Gao Shuai Yang Jiaying Liu*
Wangxuan Institute of Computer Technology, Peking University
{gaoxiang1102, williamyang, liujiaying}@pku.edu.cn

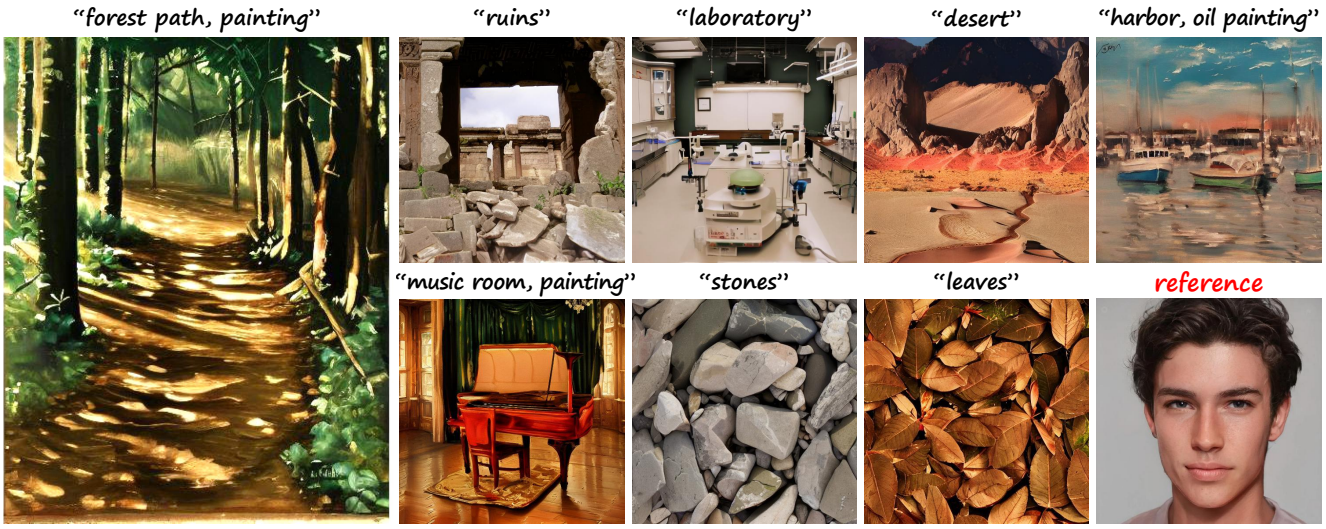

Figure 1. Taking the first image on the left as an example, what do you see at your first glance? A painting of a path through a forest (zoom in for a detailed look), or a human face (zoom out for a more global view)? Based on the off-the-shelf text-to-image diffusion model, we contribute a plug-and-play method that naturally dissolves a reference image (shown in the bottom-right corner) into arbitrary scenes described by a text prompt, providing a free lunch for synthesizing optical illusion hidden pictures using diffusion model. Better zoom in.

Abstract

*Optical illusion hidden picture is an interesting visual perceptual phenomenon where an image is cleverly integrated into another picture. Established on the off-the-shelf text-to-image (T2I) diffusion model, we propose a novel text-guided image-to-image (I2I) translation framework dubbed as **P**hase-**T**ransferred **Diffusion** Model (PTDiffusion) for hidden art syntheses, which harmoniously embeds an input reference image into arbitrary scenes described by the text prompts. At the heart of our method is a plug-and-play phase transfer mechanism that dynamically and progressively transplants diffusion features' phase spectrum from the denoising process to reconstruct the reference image into the one to sample the generated illusion image, realizing deep fusion of the reference structural information and the textual semantic information. Furthermore, we propose asynchronous phase transfer to enable flexible control over the degree of hidden content discernability. Our method bypasses any model training and fine-tuning process, all while substantially outperforming related methods in image quality, text fidelity, visual discernibility, and contextual naturalness for illusion picture synthesis, as demonstrated by extensive qualitative and quantitative experiments. Our project is publically available at this web page.*

1. Introduction

As a special form of artistic design, optical illusion hidden picture exploits human visual system's tendency to perceive patterns, shapes, and colors to conceal a secondary image within the intricate details of a primary image. It has wide applications across various fields, such as enhancing aesthetic appeal in fashion design, creating amusing content in digital entertainment, attracting attention in marketing and

* Corresponding author.

advertising, improving observation skills in children education, and visual discernment diagnosis in medical treatment.

Computationally generating optical illusions is a long-standing challenging task in computer vision and computer graphics. Early methods focus on exploiting how human brains process visual stimuli to generate elementary visual illusions, such as geometric illusion [5], color illusion [13], motion illusion [7], and viewing distance illusion [24].

More relevant to image processing, Chu *et al.* [2] propose a re-texturing pipeline to synthesize camouflage images, *i.e.*, conceal a foreground image patch into the textures of a background image. Zhang *et al.* [38] design a series of optimization functions to synthesize camouflage images from a style transfer perspective [9]. Lamdouar *et al.* [18] propose to employ StyleGAN-based generative model [16] to synthesize camouflage images in a data-driven manner.

Since diffusion models [15] revolutionizing the field of generative AI, tremendous attention has been focused on various diffusion-based AIGC applications, among which there are also explorations in illusion picture synthesis. For example, DiffQRCoder [19] and Text2QR [36] leverage ControlNet [37] to integrate scannable QR codes into aesthetic pictures. Diffusion Illusions [1] employs T2I diffusion model and score distillation sampling [27, 35] to synthesize images with overlay illusions. Visual Anagrams [10] merges noises estimated from different views to generate multi-view optical illusions, realizing image appearance change under a certain pixel permutation such as image flip, image rotation, or jigsaw rearrangement.

In this paper, we pioneer generating optical illusion hidden pictures (we will use "illusion pictures" as an abbreviation in the following) from the perspective of text-guided I2I translation, *i.e.*, translating an input reference image into an illusion picture that complies with the text prompt in semantic content while manifesting structural visual cues of the reference image. Our goal differs from the aforementioned optical illusion methods in three aspects: (i) different from camouflage image generation [2, 18, 38] that overemphasizes content concealment, we pursue visual discernibility of both target semantic content and hidden visual cues; (ii) unlike synthesizing camouflage image that conceals content into the texture of an existing background image, we expect generating background elements as per the text description; (iii) we do not aim at producing transformation-based (flip, rotation, *etc.*) optical illusions like Visual Anagrams [10], but rather focus on seamlessly dissolving a reference image into arbitrary scenes. By contrast, our goal is more methodologically relevant to text-guided I2I [21, 25, 34] and controllable T2I [22, 37, 39] methods. However, since these methods over-bind I2I correlation by explicitly enforcing feature consistency [21, 25, 34] or directly training a control network [22, 37], they are less suitable for I2I translation with large semantic deviation (*i.e.,* suffer from structure-semantic conflict issue), and thus tend to generate contextually unnatural results when applied to synthesize illusion pictures. This enlightens us to explore a disentangled image structure representation to relax I2I correlation binding, as well as an appropriate manner to deeply fuse image structure and semantic information along the sampling process.

Drawing inspiration from digital signal processing that the phase spectrum of an image determines its structural composition, we propose to leverage diffusion features' phase to disentangle image structure and accordingly propose PTDiffusion, a concise and elegant method based on T2I diffusion model that realizes smooth blending of the reference image's structural cues and the text-indicated semantic content in the Latent Diffusion Model (LDM) [29] feature space, producing visually appealing illusion pictures in a plug-and-play manner. Specifically, we employ DDIM inversion [32] to construct guidance features along a reference image reconstruction trajectory, and progressively transplant the phase of the guidance features into the corresponding features along the text-guided sampling trajectory, such that structural cues of the input reference image are smoothly penetrated into the sampling process of the target image, yielding generation results exhibiting harmonious illusion effects. Besides, we further propose asynchronous phase transfer to flexibly control structural penetration strength, endowing our method with controllability to hidden content discernability. Our method is free from training, fine-tuning, and online optimization, all while demonstrating noticeable strengths in illusion picture synthesis. The contributions are summarized as follows:

- We pioneer generating optical illusion hidden pictures from the perspective of text-guided I2I translation.
- We propose a concise and elegant method that realizes deep fusion of image structure and text semantics via dynamic phase manipulation in the LDM feature space, producing contextually harmonious illusion pictures.
- We propose asynchronous phase transfer to enable flexible control over the degree of hidden image discernibility.
- Our method dispenses with any training and optimization process, providing a free lunch for synthesizing illusion pictures using off-the-shelf T2I diffusion model.

2. Related work

Text-guided image generation. Since the advent of DDPM [15], diffusion model has soon surpassed GAN [11] on image synthesis [3] and has subsequently been accelerated by DDIM [32] and extended to conditional image generation paradigm by Palette [30]. After large-scale T2I diffusion models [23, 28, 31] remarkably boosting AIGC industry, LDM [29] contributes a classical T2I framework with dramatically lowered computational overhead by transferring DDPM from high-dimensional pixel space into low-dimensional feature space, which inspires subsequent T2I

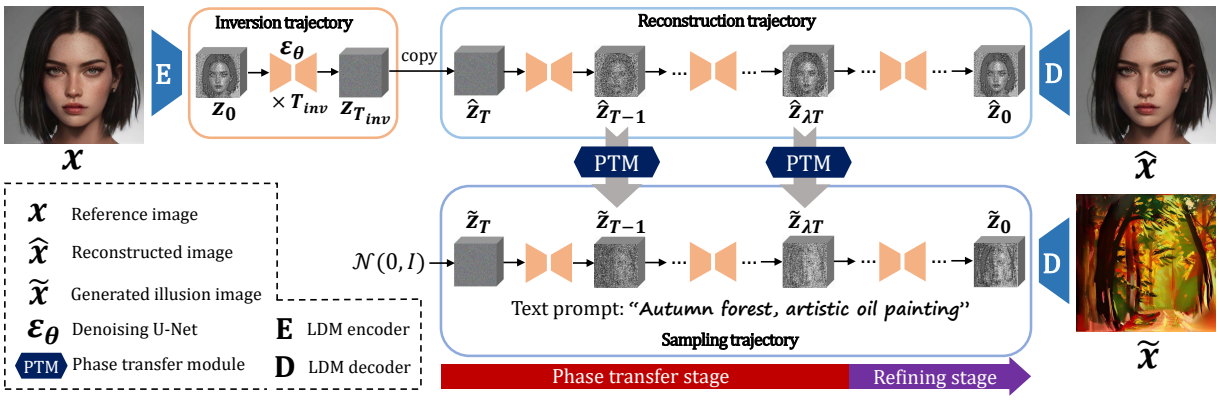

Figure 2. Overview of PTDiffusion. Built upon the pre-trained Latent Diffusion Model (LDM), PTDiffusion is composed of three diffusion trajectories. The inversion trajectory inverts the reference image into the LDM Gaussian noise space. The reconstruction trajectory recovers the reference image from the inverted noise embedding. The sampling trajectory samples the final illusion image from random noise guided by the text prompt. The reconstruction and sampling trajectory are bridged by our proposed phase transfer module, which dynamically transplants diffusion features' phase spectrum to smoothly blend source image structure with textual semantics in the LDM feature space.

models [6, 26] scaling up to larger capacity. To add controllability to T2I synthesis, ControlNet [37] and T2I-Adapter [22] train a control network of the LDM conditioned on certain image priors (edges, depth maps, *etc.*), implicitly realizing reference-image-based structural control to the generated images. However, the over-constraining on object contours and shapes of these methods limit their applicability to synthesizing illusion pictures which emphasize harmonious blending of source image structure and target semantics.

Text-guided I2I translation. Our method closely relates to text-guided I2I translation. SDEdit [20] translates a source image by noising it to an intermediate step followed by text-guided denoising. Attention-modulation-based methods such as Null-text inversion [21] and pix2pix-zero [25] correlate source and generated image by enforcing consistency of cross-attention maps [12]. Textual-inversion-based methods like Imagic [17] and Prompt Tuning Inversion [4] preserve source image visual information via learnable text embedding. Optimization-free methods such as PAP [34] and FBSDiff [8] maintain source image structure through dynamic feature modulation during the reverse denoising process. For illusion picture synthesis, however, these methods struggle to produce contextually natural results with both faithful textual semantics and discernable hidden structure due to the overly bound I2I correlation.

3. Phase-Transferred Diffusion Model

3.1. Overall architecture

As Fig. 2 shows, our model builds on the off-the-self LDM [29], and is comprised of an inversion trajectory ($z_0 \rightarrow z_{T_{inv}}$), a reconstruction trajectory ($z_{T_{inv}} = \hat{z}_T \rightarrow \hat{z}_0 \approx z_0$), and a sampling trajectory ($\tilde{z}_T \rightarrow \tilde{z}_0$). Based on the initial feature $z_0 = E(x)$ extracted from the reference image x by the LDM encoder E, the inversion trajectory adopts a T_{inv}-step DDIM inversion [32] to project z_0 into a Gaussian noise $z_{T_{inv}}$ conditioned on the null-text embedding $v_\emptyset$:

$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}} f_\theta(z_t, t, v_\emptyset) + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(z_t, t, v_\emptyset), \quad (1)$$

$$f_\theta(z_t, t, v_\emptyset) = (z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, v_\emptyset))/\sqrt{\bar{\alpha}_t}, \quad (2)$$

where $\{\bar{\alpha}_t\}$ are pre-defined DDPM schedule parameters [15], ϵ_θ is the LDM denoising U-Net, $f_\theta(z_t, t, v_\emptyset)$ is an approximation of z_0 estimated from z_t. The reconstruction trajectory applies a T-step DDIM sampling to reconstruct $\hat{z}_0 \approx z_0$ from the inverted Gaussian noise $\hat{z}_T = z_{T_{inv}}$, conditioned on the same null-text embedding $v_\emptyset$:s

$$\hat{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_\theta(\hat{z}_t, t, v_\emptyset) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\hat{z}_t, t, v_\emptyset), \quad (3)$$

where $f_\theta(\hat{z}_t, t, v_\emptyset)$ is similar to Eq. (2). The sampling trajectory applies a T-step DDIM sampling to generate $\tilde{z}_0$ from a randomly initialized Gaussian noise $\tilde{z}_T \sim \mathcal{N}(0, I)$ conditioned on the text embedding v of the target text prompt. To amplify the influence of text guidance, we exploit classifier-free guidance technique [14] by linearly combining the conditional (target text) and unconditional (null text) noise estimation with a guidance scale ω:

$$\tilde{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_\theta(\tilde{z}_t, t, v, v_\emptyset) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\tilde{z}_t, t, v, v_\emptyset), \quad (4)$$

$$f_\theta(\tilde{z}_t, t, v, v_\emptyset) = (\tilde{z}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\tilde{z}_t, t, v, v_\emptyset))/\sqrt{\bar{\alpha}_t}, \quad (5)$$

$$\epsilon_\theta(\tilde{z}_t, t, v, v_\emptyset) = \omega \cdot \epsilon_\theta(\tilde{z}_t, t, v) + (1-\omega) \cdot \epsilon_\theta(\tilde{z}_t, t, v_\emptyset). \quad (6)$$

To dissolve x into $\tilde{x}$, we propose phase transfer module (PTM) which dynamically blends the structural information of $\hat{z}_t$ into $\tilde{z}_t$ along the two parallel denoising trajectories. We apply per-step structural penetration realized by PTM only in the early part of the sampling trajectory (which we

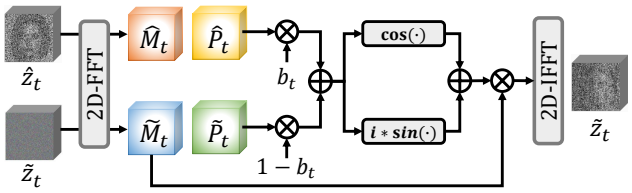

Figure 3. Illustration of the phase transfer module (PTM). To transfer the phase of $\hat{z}_t$ into $\tilde{z}_t$, we apply 2D FFT to decompose their magnitude $\hat{M}_t$, $\tilde{M}_t$ and phase $\hat{P}_t$, $\tilde{P}_t$, linearly fuse their phase with a blending coefficient b_t, and recombine the fused phase with $\tilde{M}_t$. Finally, the manipulated FFT feature is converted back to the spatial domain via 2D IFFT to form the phase-transferred $\tilde{z}_t$.

call phase transfer stage) while leaving the rear part (which we call refining stage) totally unconstrained to guarantee high-quality image synthesis. The two stages are separated by the time step λT, where λ denotes the proportion of the refining stage to the entire sampling trajectory. The final sampling result $\tilde{z}_0$ is transformed to the generated illusion picture via the LDM decoder D, *i.e.*, $\tilde{x} = D(\tilde{z}_0)$.

3.2. Phase transfer module

As the kernel ingredient of our method, the PTM is illustrated in Fig. 3. To transfer the phase of $\hat{z}_t$ into the corresponding feature $\tilde{z}_t$, we firstly utilize 2D Fast Fourier Transform (FFT) to extract their magnitude and phase:

$$\hat{R}_t + i\hat{I}_t = FFT(\hat{z}_t), \tilde{R}_t + i\tilde{I}_t = FFT(\tilde{z}_t), \quad (7)$$

$$\hat{M}_t = \sqrt{\hat{R}_t^2 + \hat{I}_t^2}, \tilde{M}_t = \sqrt{\tilde{R}_t^2 + \tilde{I}_t^2}, \quad (8)$$

$$\hat{P}_t = arctan(\hat{I}_t / \hat{R}_t), \tilde{P}_t = arctan(\tilde{I}_t / \tilde{R}_t), \quad (9)$$

where i denotes the imaginary unit, *i.e.*, $i^2 = -1$. $\hat{R}_t$ and $\tilde{R}_t$ are the real part of the 2D FFT spectrum of $\hat{z}_t$ and $\tilde{z}_t$ respectively, $\hat{I}_t$ and $\tilde{I}_t$ are the corresponding imaginary part. $\hat{M}_t$ and $\tilde{M}_t$ are the magnitude spectrum of $\hat{z}_t$ and $\tilde{z}_t$ respectively, $\hat{P}_t$ and $\tilde{P}_t$ are the corresponding phase spectrum. Then, the extracted phase spectrum $\hat{P}_t$ and $\tilde{P}_t$ are linearly blended with a time-dependent blending coefficient b_t, yielding the fused phase spectrum P_t^{fuse} as follows:

$$P_t^{fuse} = b_t \times \hat{P}_t + (1 - b_t) \times \tilde{P}_t. \quad (10)$$

The fused phase is recombined with the original magnitude $\tilde{M}_t$ before transforming back to the spatial domain with 2D IFFT, finally resulting in the structurally penetrated $\tilde{z}_t$:

$$\tilde{z}_t = IFFT(\tilde{M}_t \times (\cos(P_t^{fuse}) + i \times \sin(P_t^{fuse}))). \quad (11)$$

Since the structural information of the guidance features $\{\hat{z}_t\}$ is becoming increasingly prominent as the denoising process proceeds, direct phase transfer in the later denoising steps is prone to harm contextual naturalness of the final result due to excessive structural penetration. Thus, we design

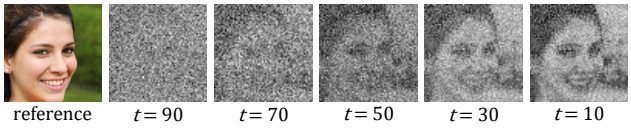

Figure 4. Visualization of the guidance features $\{\hat{z}_t\}$ along the 100-step reconstruction trajectory. The structural information of $\hat{z}_t$ becomes increasingly distinct as the denoising proceeds.

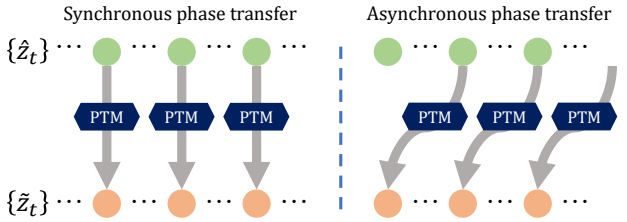

Figure 5. Illustration of the asynchronous phase transfer which transfers phase across diffusion features at different time steps.

a decayed phase blending schedule which gradually decays the blending coefficient $\{b_t\}$ in the later denoising steps of the phase transfer stage, which we formulate as below:

$$b_t = \begin{cases} 1, & if \ \ \tau T \le t \le T \\ 1 - \sqrt{\dfrac{\tau T - t}{\tau T - \lambda T}}, & if \ \ \lambda T \le t < \tau T \end{cases} \quad (12)$$

where $\lambda \le \tau \le 1$. It means that we apply direct phase replacement in the early part of the phase transfer stage where the guidance features $\hat{z}_t$ are structurally less distinct, while gradually decaying phase transfer intensity in the later section of the phase transfer stage as $\hat{z}_t$ becomes structurally more and more prominent. By decaying phase transfer intensity, the visual quality and contextual naturalness of the generated illusion picture can be noticeably improved.

3.3. Asynchronous phase transfer

Since sufficient steps in the refining stage is of crucial importance to ensure image quality, we fix λ, namely the length of the phase transfer stage, and propose asynchronous phase transfer to realize controllable structural penetration strength within the fixed-length phase transfer stage. As Fig. 4 displays, later $\hat{z}_t$ in the reconstruction trajectory is structurally more prominent than its earlier counterpart. This inspires us to transfer phase from later $\hat{z}_t$ into earlier $\tilde{z}_t$ to enhance structural penetration, namely the so-called asynchronous phase transfer as illustrated in Fig. 5.

To this end, we design a concise and elegant solution that implements asynchronous phase transfer based on simple synchronous denoising, *i.e.*, parallel denoising along the reconstruction and sampling trajectory. Specifically, given an async distance d, we firstly leverage $\hat{z}_t$ to estimate its future counterpart $\hat{z}_{t-d}$ that is d steps ahead of itself in the recon-

Algorithm 1 Complete algorithm of PTDiffusion

Input: Reference image x, target text v, time steps T and T_{inv}, blending coefficients $\{b_t\}$, async distance d.
Output: The generated illusion picture $\tilde{x}$.

Extract the initial latent feature $z_0 = E(x)$
for $t = 0$ to $T_{inv} - 1$ **do**
  Compute z_{t+1} from z_t via Eq. (1)
end for{Inversion trajectory}
Initialize $\hat{z}_T = z_{T_{inv}}$, $\tilde{z}_T \sim \mathcal{N}(0, I)$
for $t = T$ to λT **do**
  Compute $\hat{z}_{t-1}$ from $\hat{z}_t$ via Eq. (3)
  Compute $\tilde{z}_{t-1}$ from $\tilde{z}_t$ via Eq. (4)
  $\tilde{z}_{t-1} = APTM(\hat{z}_{t-1}, \tilde{z}_{t-1}, b_{t-1}, d)$ as Eq. (13)
end for{Phase transfer stage}
for $t = \lambda T - 1$ to 1 **do**
  Compute $\tilde{z}_{t-1}$ from $\tilde{z}_t$ via Eq. (4)
end for{Refine stage}
Compute the illusion picture $\tilde{x} = D(\tilde{z}_0)$

struction trajectory, then transfer the phase spectrum of the estimated $\hat{z}_{t-d}$ into $\tilde{z}_t$. Let APTM denote the asynchronous phase transfer module, we formulate it as follows:

$$\tilde{z}_t = APTM(\hat{z}_t, \tilde{z}_t, b_t, d) = PTM(\hat{z}^*_{t-d}, \tilde{z}_t, b_t), \quad (13)$$

where $\hat{z}^*_{t-d}$ is a pre-estimation of $\hat{z}_{t-d}$ at time step t:

$$\hat{z}^*_{t-d} = \sqrt{\bar{\alpha}_{t-d}} f_\theta(\hat{z}_t, t, v_\emptyset) + \sqrt{1 - \bar{\alpha}_{t-d}} \epsilon_\theta(\hat{z}_t, t, v_\emptyset). \quad (14)$$

Similar to Eq. (2), $f_\theta(\hat{z}_t, t, v_\emptyset)$ denotes an approximate estimation of $\hat{z}_0$ predicted by the current $\hat{z}_t$. Note that the async distance d in Eq. (13) can also be a negative value to allow for weakened structural penetration strength.

3.4. Implementation details

We use pre-trained SD v1.5 as the backbone diffusion model and set the classifier-free guidance scale ω=7.5. To ensure inversion accuracy, we apply 1000-step DDIM inversion for the inversion trajectory, *i.e.*, T_{inv}=1000. We apply 100-step DDIM sampling for both the reconstruction and sampling trajectory to save inference time, *i.e.*, T=100. During sampling, we allocate 60% denoising steps to the phase transfer stage and the remaining 40% steps to the refining stage, *i.e.*, λ=0.4. We perform direct phase replacement in the early $2/3$ section of the phase transfer stage while performing decayed phase transfer in the later $1/3$ section by setting τ=0.6 in Eq. (12). The async distance d in Eq. (13) is manually tunable (recommended in the range of [-10, 10]) for flexible control over the structural penetration strength, with a default value of 0. The complete algorithm of PTDiffusion is summarized in Alg. 1.

4. Experiment

4.1. Qualitative results

Example results of our PTDiffusion in generating illusion pictures are displayed in Fig. 6. Our method harmoniously dissolves a reference image into arbitrary scenes described by a text prompt. Apart from real pictures as input images, our method also supports synthetic ones, *e.g.*, integrating a binary text image into the scene of the corresponding semantics to generate contextual text images. It also shows that our method is capable of producing visually appealing illusion pictures of both realistic and artistic domain.

As demonstrated in Fig. 7, our method allows to sample diverse illusion pictures simply by varying the initialized Gaussian noise $\tilde{z}_T$, while existing advanced text-guided I2I methods [21, 25, 34] do not possess such diversity property.

In Fig. 9, we visually compare our method with related text-guided I2I methods including Null-text inversion (NTI) [21], Prompt-tuning inversion (PTI) [4], PAP [34], FBSDiff [8], and SDEdit [20], as well as controllable T2I method represented by ControlNet [37]. We test different denoising strengths for SDEdit, as well as different conditioning modes (Canny edge and depth map) and control weights for ControlNet for comprehensive evaluation of their performance in producing illusion effects. For the remaining methods, we present results with best illusion effect after hyperparameter tuning. NTI produces results with weak target semantics due to the overly bound I2I correlation caused by the attention map consistency constraint. Results of PAP are more aligned to the target semantics but still fall short in text fidelity. It also has the structural overbinding issue due to directly reusing reference features during sampling. Results of PTI and FBSDiff manifest a certain degree of structure-semantic blending, but still underperform in contextual naturalness reflected by visually unpleasant artifacts. Moreover, they suffer from issue of less prominent hidden content, and fail to suit binary text reference images. Results of SDEdit show its difficulty in balancing structure-semantic trade-off, *i.e.*, large denoising strength produces results with overwhelmed structural cues while small one is insufficient to translation input image to the target semantics. Likewise, ControlNet suffers from the same challenge to balance source structure and target semantics when tuning the control weight. By contrast, our PTDiffusion is the only one among the compared methods that realizes harmonious structure-semantic blending, producing visually appealing illusion pictures manifesting both precise textual semantics and clearly discernible hidden content.

Qualitative ablation studies shown in Fig. 8 demonstrate that both the phase transfer decay and the refining stage contribute to improving contextual naturalness of the structure-semantic blending. The absence of each of them leads to overly penetrated reference structure, degrading visual har-

Figure 6. Example results of PTDiffusion in generating optical illusion hidden pictures. Better viewed with zoom-in.

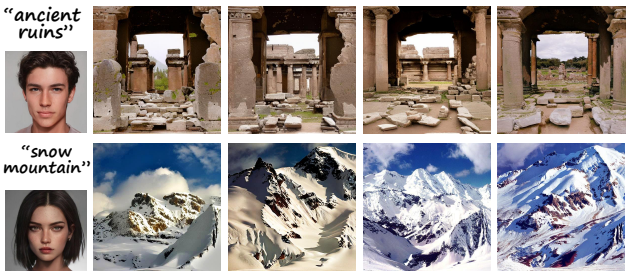

Figure 7. Our method allows to sample diverse illusion pictures.

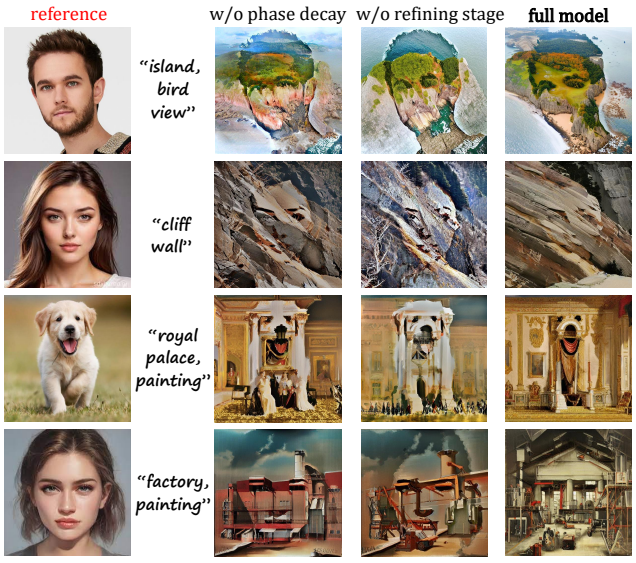

Figure 8. Ablation study about the phase transfer decay and the refining stage of our method. Better viewed with zoom-in.

mony of the generated illusion pictures.

To verify the necessity of DDIM inversion, we have experimented with using the forward diffusion process (*i.e.*, adding noise to z_0 according to $\hat{z}_t \sim \mathcal{N}(\hat{z}_t; \sqrt{\bar{\alpha}_t} z_0, (1 - \bar{\alpha}_t)\mathcal{I})$) to build the guiding trajectory $\{\hat{z}_t\}$, which we term the model w/o inversion. As Fig. 10 shows, our results w/ inversion are superior to that w/o inversion in both visual quality and contextual naturalness. This could be due to that the forward diffusion introduces randomness (Gaussian noise) to the guidance features, yielding unstable guiding trajectory with feature phase irregularly perturbed at each time step, while the guiding trajectory built with DDIM inversion is totally deterministic, providing more stable phase spectra to be transferred along the sampling process.

Fig. 11 qualitatively demonstrates the effectiveness of our proposed asynchronous phase transfer in controlling the degree of hidden image discernibility. Increasing async distance d enhances hidden content visual prominence by transferring phase from later guidance features into earlier sampling features, while reducing d weakens hidden content discernibility by conversely transferring phase from earlier guidance features into later sampling features.

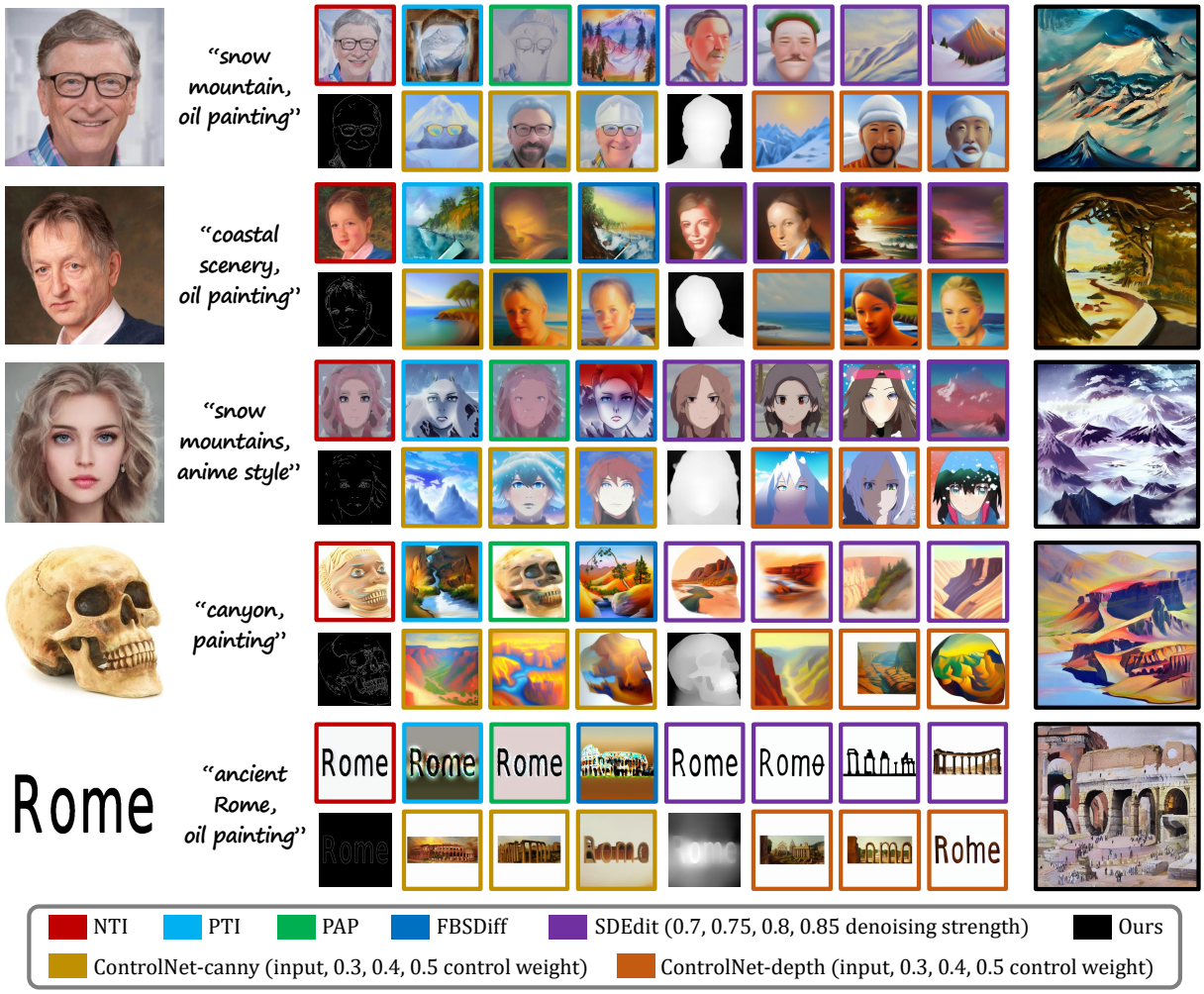

Figure 9. Comparison to related text-guided I2I and controllable T2I methods on generating illusion pictures. Better to zoom in.

Figure 10. Comparison between using inversion and using forward diffusion (w/o inversion) to construct the guiding trajectory.

Table 1. Quantitative evaluations of different text-guided I2I methods for illusion picture synthesis. Results of ControlNet are obtained under the depth condition with 0.4 control weight. Results of SDEdit are evaluated under 0.8 denoising strength.

Method	Aesthetic Score (↑)	CLIP Score (↑)	LPIPS (↑)
NTI [21]	6.24	0.23	0.21
PTI [4]	6.18	0.28	0.52
PAP [34]	6.09	0.25	0.40
FBSDiff [8]	5.96	0.29	0.56
SDEdit [20]	6.10	0.29	0.49
ControlNet [37]	6.05	0.26	0.47
PTDiffusion (ours)	**6.37**	**0.31**	**0.64**

4.2. Quantitative results

For quantitative evaluations, we measure image visual quality using Aesthetic Score (↑) predicted by the pre-trained LAION Aesthetics Predictor V2 model, measure text fidelity using CLIP Score (↑), *i.e.*, the image-text cosine similarity. To measure model's ability to modify source image appearance, we evaluate LPIPS (↑) between the reference and the generated image pair. We test on 80 reference images with each one paired with at least 3 text prompts.

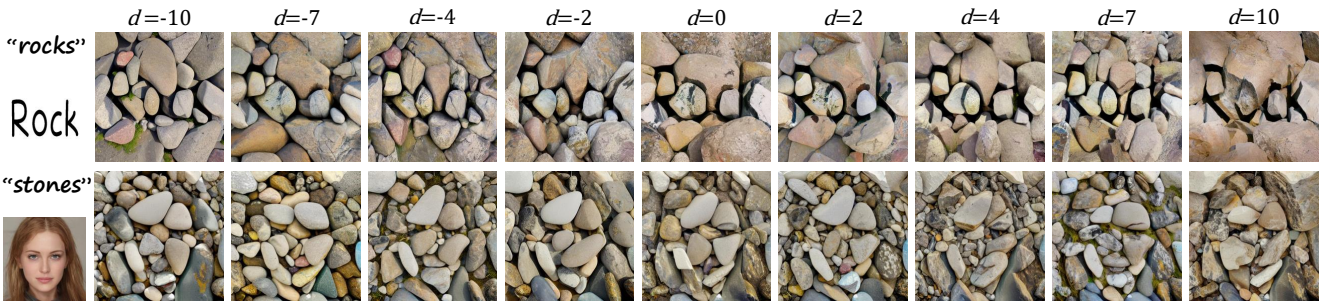

Figure 11. Demonstration of the hidden content discernibility control of our method realized by varying the async distance parameter d.

Table 2. Quantitative ablation study *w.r.t* phase intensity decay, refining stage, and DDIM inversion.

Model	Aesthetic Score (↑)	CLIP Score (↑)
w/o phase decay	6.24	0.28
w/o refining stage	5.95	0.24
w/o inversion	6.12	0.25
full model	**6.37**	**0.31**

Table 3. Quantitative study of the the impact of the async distance d to structural penetration strength and text fidelity.

Async distance d	-9	-6	-3	0	3	6	9
Structure Similarity (↑)	0.880	0.893	0.902	0.908	0.912	0.917	0.926
CLIP Score (↑)	0.309	0.311	0.305	0.307	0.301	0.298	0.293

Results reported in Tab. 1 show that our method achieves leading performance for all the aforementioned metrics.

We quantitatively ablate the influence of the phase transfer decay, the refining stage, and the DDIM inversion *w.r.t* the image quality and text fidelity. Results displayed in Tab. 2 show that missing any one of these ingredients results in declined aesthetic score and CLIP similarity score, which is basically in line with the qualitative results displayed in Fig. 8 and Fig. 10. Moreover, the absence of the refining stage causes the most performance drop for these two metrics.

To quantitatively prove the hidden content perceptibility control ability of our proposed asynchronous phase transfer, we quantify hidden content visual prominence as the structure similarity between input and output image pair, for which we use DINO-ViT self-similarity score [33] as the metric. As reported in Tab. 3, the I2I structure similarity continuously grows with the increase of the async distance d, which tallies with the qualitative results of Fig. 11. It is also worth noting that the increase of d does not lead to drastic drop in CLIP Score, indicating that our proposed asynchronous phase transfer promotes hidden content discernibility without noticeably sacrificing text fidelity.

For aesthetic assessment of the generated optical illusion hidden pictures, we resort to user study for subjective evaluation. Based on the unique visual characteristics of illusion pictures, we invite 16 participants to score the generation results of different methods on a scale of 1-10 from the following two perspectives: (i) contextual naturalness, *i.e.*, to what extent the reference image is naturally and reasonably blended into the textual scene content; (ii) illusion balance, *i.e.*, well-balanced visual discernibility of the textual scene content and the original hidden content, rather than visual prominence of only one side. The average user ratings of all the compared methods in Tab. 1 are reported in Fig. 12, our method outscores other approaches by a large margin in both two dimensions, subjectively indicating significant advantage of our PTDiffusion in illusion picture synthesis.

Figure 12. Average user ratings of different methods.

5. Conclusion

This paper pioneers generating optical illusion hidden pictures from the perspective of text-guided I2I translation, *i.e.*, translating a reference image into an illusion picture that is faithful to the text prompt in semantic content while manifesting perceptible structural cues of the reference image. To this end, we propose PTDiffusion, a concise and novel method capable of synthesizing contextually harmonious illusion pictures based on the off-the-shelf T2I diffusion model. At the core of our method is a plug-and-play phase transfer module that smoothly fuses the reference structural information with the textual semantic information via progressive phase transfer between the latent diffusion features. We further propose asynchronous phase transfer to enable flexible control over hidden content discernibility. Though dispensing with model training and fine-tuning, our method shows significant advantages in hidden art synthesis.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 62332010 and 62471009, in part by CCF-Tencent Rhino-Bird Open Research Fund, and in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

[1] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. In *Proceedings of the ACM SIGGRAPH*, pages 1–11, 2024. 2

[2] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. In *Proceedings of the ACM SIGGRAPH*, pages 1–8. 2010. 2

[3] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 8780–8794, 2021. 2

[4] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7430–7440, 2023. 3, 5, 7

[5] Werner Ehm. A variational approach to geometric-optical illusions modeling. *Proceedings of Fechner Day*, 27(1):41–46, 2011. 2

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*, 2024. 3

[7] William T Freeman, Edward H Adelson, and David J Heeger. Motion without movement. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 27–30, 1991. 2

[8] Xiang Gao and Jiaying Liu. Fbsdiff: Plug-and-play frequency band substitution of diffusion features for highly controllable text-driven image translation. In *Proceedings of the ACM International Conference on Multimedia*, pages 4101–4109, 2024. 3, 5, 7

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 2

[10] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *Proceedings of The International Conference on Learning Representations*, 2023. 3

[13] Elad Hirsch and Ayellet Tal. Color visual illusions: a statistics-based computational model. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 9447–9458, 2020. 2

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 6840–6851, 2020. 2, 3

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[17] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[18] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 832–842, 2023. 2

[19] Jia-Wei Liao, Winston Wang, Tzu-Sian Wang, Li-Xuan Peng, Ju-Hsuan Weng, Cheng-Fu Chou, and Jun-Cheng Chen. DiffQRCoder: Diffusion-based aesthetic qr code generation with scanning robustness guided iterative refinement. *arXiv preprint arXiv:2409.06355*, 2024. 2

[20] 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. *arXiv preprint arXiv:2108.01073*, 2021. 3, 5, 7

[21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 3, 5, 7

[22] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2, 3

[23] 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 *Proceedings of the International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[24] Aude Oliva, Antonio Torralba, and Philippe G Schyns. Hybrid images. *ACM Transactions on Graphics*, 25(3):527–532, 2006. 2

[25] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image

translation. In *Proceedings of the ACM SIGGRAPH*, pages 1–11, 2023. 2, 3, 5

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[27] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[28] 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. 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[30] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Proceedings of the ACM SIGGRAPH*, pages 1–10, 2022. 2

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Raphael Gontijo-Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 36479–36494, 2022. 2

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[33] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 8

[34] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2, 3, 5, 7

[35] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2D diffusion models for 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[36] Guangyang Wu, Xiaohong Liu, Jun Jia, Xuehao Cui, and Guangtao Zhai. Text2QR: Harmonizing aesthetic customization and scanning robustness for text-guided QR code generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2024. 2

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 5, 7

[38] Qing Zhang, Gelin Yin, Yongwei Nie, and Wei-Shi Zheng. Deep camouflage images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12845–12852, 2020. 2

[39] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-ControlNet: all-in-one control to text-to-image diffusion models. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 11127–11150, 2023. 2