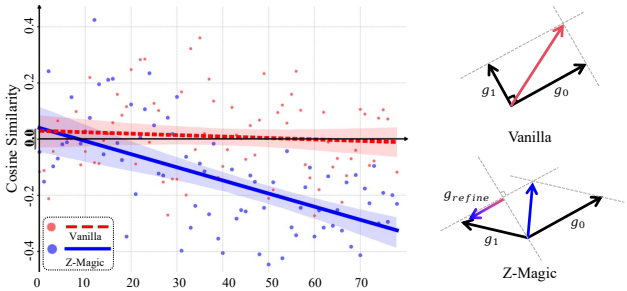
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Z-Magic: Zero-shot Multiple Attributes Guided Image Creator

Yingying Deng, Xiangyu He[1],Fan Tang[✉,2], Weiming Dong[1]
[1] MAIS, Institute of Automation, Chinese Academy of Sciences
[2] Institute of Computing Technology, Chinese Academy of Sciences
dyy15@outlook.com, tfan.108@gmail.com, weiming.dong@ia.ac.cn

Abstract

The customization of multiple attributes has gained popularity with the rising demand for personalized content creation. Despite promising empirical results, the contextual coherence between different attributes has been largely overlooked. In this paper, we argue that subsequent attributes should follow the multivariable conditional distribution introduced by former attribute creation. In light of this, we reformulate multi-attribute creation from a conditional probability theory perspective and tackle the challenging zero-shot setting. By explicitly modeling the dependencies between attributes, we further enhance the coherence of generated images across diverse attribute combinations. Furthermore, we identify connections between multi-attribute customization and multi-task learning, effectively addressing the high computing cost encountered in multi-attribute synthesis. Extensive experiments demonstrate that Z-Magic outperforms existing models in zero-shot image generation, with broad implications for AI-driven design and creative applications.



(a) Gradient Cosine Similarity (b) Gradient Direction

Figure 1. We visualize the cosine similarity between gradients introduced by two conditions (i.e., landmark g_0 and face ID g_1) during the diffusion process of face synthesis. In the vanilla setting, each condition is applied separately without considering contextual coherence, causing the gradient directions to be nearly orthogonal, as random vectors in high-dimensional space are often orthogonal. In contrast, our method adjusts each condition based on the preceding ones, resulting in an obtuse-angled optimization direction where later conditions refine the gradients of earlier ones. In other words, our approach has two degrees of freedom to construct the gradient direction: angle and length.

1. Introduction

With the rapid advancement of generative AI, particularly in the widespread adoption of image generation applications, the paradigm of controlled image creation has gained considerable attention from both industry and the image synthesis community [34, 36–38]. In this context, guiding the image creation process through multimodal conditions, such as text prompts and reference images, becomes more popular.

In reality, conditional dependence is a fundamental property in multi-attribute guided creation. For example, in fashion design, the selection of a color scheme is often influenced by the style or cut of the clothing, and vice versa; similarly, in face synthesis, the presence of gender implies facial features, and this contextual information should guide subsequent attribute choices. These examples intuitively illustrate that attributes do not exist in isolation but are mutually informative and conditionally dependent, forming a naturally coherent whole. While existing diffusion-based approaches have achieved impressive results in conditional generation [3–5, 15, 19, 22, 44, 48], the underlying structure that inherently links multiple attributes into a cohesive creation has been rarely discussed. Besides, for the sake of practical simplicity, previous methods often assume independence among different attributes, which limits their capacity to model inter-attribute relationships.

In the score-based conditional diffusion model, conditioning is facilitated through a gradient-based approach, formulated as $g_{\mathbf{c}} = \nabla_{\mathbf{x}_t} \log p(\mathbf{c}|\mathbf{x}_t)$, where the gradient $g_{\mathbf{c}}$ represents the conditioning mechanism. When each attribute is generated under the assumption of independence, the gradients associated with distinct attributes exhibit near-

✉Corresponding author: Fan Tang.

zero cosine similarity. As shown in Figure 1(a), for any two conditional gradients, $g_{\mathbf{c}_0}$ and $g_{\mathbf{c}_1}$, during the diffusion process, we observe that $\mathbb{E}\left[\frac{g_{\mathbf{c}_0}^T g_{\mathbf{c}_1}}{\|g_{\mathbf{c}_0}\|\|g_{\mathbf{c}_1}\|}\right] \approx 0$. This behavior aligns with the properties of random high-dimensional vectors, which are typically nearly orthogonal in high-dimensional spaces . This result is somewhat surprising, as these conditions apply to the same image, and ideally, they should exhibit a degree of contextual correlation to support coherent and harmonious synthesis.

In light of this, we present a novel approach that explicitly models conditional dependencies, ensuring each attribute is generated in the context of previously selected ones. Specifically, our method employs a conditional distribution framework, where each subsequent attribute is derived from a multivariable conditional distribution informed by the attributes generated earlier, aligning with the joint distribution $p(\mathbf{c}_0, \mathbf{c}_1|\mathbf{x}_t)$. Each conditionally dependent gradient progressively refines image generation. Initially, we apply $g_0 = \nabla \mathbf{x}_t \log p(\mathbf{c}_0|\mathbf{x}_t)$ to capture the influence of the first attribute, followed by $g_1 = \nabla \mathbf{x}_t \log p(\mathbf{c}_1|\mathbf{c}_0, \mathbf{x}_t)$, which adjusts the generation to align with the contextual influence of the first attribute. As shown in Figure 1(b), this sequential conditioning introduces a gradient refinement mechanism whereby g_1 partially aligns with g_0 and corrects the vector length along the g_0 direction, enhancing attribute coherence.

Further, while models can be fine-tuned for specific attribute combinations, training-based approach is resource-intensive and lacks scalability, particularly in the face of high-dimensional attribute spaces. Thus, exploring methods for zero-shot multi-attribute synthesis has the potential to unlock new avenues for practical AI-driven creative tools. In response to these challenges, we introduce Zero-shot Multiple Attributes Guided Image Creator (Z-Magic), an approach designed to address the complex requirements of multi-attribute customization. By reformulating multi-attribute synthesis as a problem of conditional dependency across attributes, Z-Magic leverages principles from conditional probability and multi-task learning to generate coherent images without extensive attribute-specific tuning. In our experiments, this method not only reduces human labeling overhead for attribute combinations but also paves the way for harmonious customization across diverse domains.

2. Related Works

2.1. Conditional Diffusion Model

Recent advances in Denoising Diffusion Probabilistic Models (DDPMs) [18] have demonstrated exceptional capabilities in various image synthesis tasks, including image manipulation and conditional generation. Classifier-Guidance [13] introduces an additional classifier to guide the generation toward specific categories, while classifier-free guidance models [17] combine conditional and unconditional models to enhance generation quality.

Several landmark architectures have further shaped this field. GLIDE [31] employs a pre-trained CLIP model [35] for text-guided image synthesis, and Stable Diffusion [38] improves computational efficiency by performing text-conditioned denoising in latent space. ControlNet [48] incorporates a parallel U-Net architecture to enable diverse visual conditions, including landmarks, edge maps, and skeletal structures. Similarly, T2I-Adapter [30] proposes a lightweight adapter network to integrate various visual control signals.

Despite their impressive performance, these approaches demand substantial computational resources for training. In contrast, our work explores a training-free paradigm for controlling diffusion models, offering greater flexibility and broader applicability.

2.2. Multiple Attributes Customization

Multi-attribute control in image generation—encompassing aspects such as style, structure, textual descriptions, and semantic segmentation—has become a critical research focus in computer vision, particularly in tasks like style transfer and facial image synthesis. StyleGAN [21] pioneered this field by introducing a style-based generator architecture that enables precise control over visual features at multiple scales, setting new standards in high-quality face image generation.

Subsequent advances have further extended the capabilities of attribute-controlled generation. HairCLIP [45] built on the StyleGAN framework by integrating both textual and reference image conditioning. TediGAN [46] advanced semantic control by combining GAN inversion techniques with style-based generation. In diffusion models, recent research has shown significant progress: for instance, [25] achieved control over both identity and expression attributes, while [22] demonstrated the complementary strengths of diffusion models and GANs for text and visually guided face generation. Additionally, [19] introduced a multi-modal face control framework using pre-trained diffusion models. Various style transfer techniques [9–12, 15] have also been developed to jointly address style and content conditions.

While these methods deliver impressive results, they often face limitations in scalability for additional control conditions. To address this, we propose a unified paradigm that enables the seamless integration of multiple attribute control signals. Our framework provides a more flexible and generalizable approach to multi-attribute-controlled image generation.

In high-dimensional spaces, two random vectors are almost always nearly orthogonal [39].

3. Approach

3.1. Preliminaries

Score-based diffusion models [20, 40–42], as a unifying framework, consolidate previous methodologies in both score-based generative modeling and diffusion probabilistic modeling. These models describe the forward diffusion process $\{\mathbf{x}(t)\}_{t=0}^{T}$ as a stochastic differential equation (SDE). Given this forward SDE, the reverse process can be formulated by solving the corresponding reverse-time SDE in the context of controlled generation [2, 42]:

$$d\mathbf{x} = \Big\{\mathbf{f}(\mathbf{x},t) - g(t)^2\big[\nabla_{\mathbf{x}}\log p_t(\mathbf{x}) + \nabla_{\mathbf{x}}\log p_t(\mathbf{c}|\mathbf{x})\big]\Big\}dt + g(t)d\bar{\mathbf{w}}, \tag{1}$$

where $\mathbf{f}(\mathbf{x},t)$ denotes the drift coefficient, and $\bar{\mathbf{w}}$ represents the standard Wiener process for reverse-time dynamics. Additionally, [42] proposes a time-dependent score-based model, $\mathbf{s}_{\boldsymbol{\theta}^*}$, to approximate $\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)$, thereby enabling the construction and numerical solution of the reverse-time SDE. Our subsequent discussion focuses on the discretization of these stochastic dynamics.

3.2. Multiple Attributes Guided Creation

There are numerous numerical methods for solving SDEs mentioned above, in this section, we use the ancestral sampling adopt in the popular DDPM [18] to ease the derivation . The sampling method serves as the solver to the reverse-time VP SDE [42], formulated as

$$\mathbf{x}_{t-1,\mathbf{c}} = (2-\sqrt{1-\beta_t})\mathbf{x}_t + \beta_t\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{c}) + \sqrt{\beta_t}\boldsymbol{\epsilon}.$$

Given the approximation that $\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t) \approx \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t,t)$, we have

$$\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{c}) \approx \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t,t) + \underbrace{\nabla_{\mathbf{x}_t}\log p(\mathbf{c}|\mathbf{x}_t)}_{\text{controllable term}}. \tag{2}$$

where s_{θ^*} is the pre-trained unconditional score estimator model, $\mathbf{c}$ refers to the set of conditions, such as text prompts, landmark points, and reference images. By expanding the single condition guidance $p(\mathbf{x}_t|\mathbf{c})$ to the multiple attributes guidance

$$p(\mathbf{x}_t|\mathbf{c}_1,...,\mathbf{c}_n) = \frac{p(\mathbf{c}_1,...,\mathbf{c}_n|\mathbf{x}_t)p(\mathbf{x}_t)}{p(\mathbf{c}_1,...,\mathbf{c}_n)}, \tag{3}$$

it contributes to more delicate manipulations on $\mathbf{x}_{t-1}$ with $\nabla_{\mathbf{x}_t}\log p(\mathbf{c}_1,...,\mathbf{c}_n|\mathbf{x}_t)$.

Please note that our approach is not limited to DDPMs and is also compatible with other solvers.

3.3. Conditional Independent Creation

When facing multiple attributes prior $\mathbf{c}_i$, suppose we were considering the position of the target object or the style of the image, these conditions themselves are independent, if they are measuring the target image among different dimensions. For instance, we have a cat in the middle of the image would not tell you anything about Van Gogh's art style. Formally,

$$p(\mathbf{c}_1,...,\mathbf{c}_n) = p(\mathbf{c}_1)p(\mathbf{c}_2)...p(\mathbf{c}_n) = \prod_{i=1}^{n} p(\mathbf{c}_i), \tag{4}$$

Following the above derivation, it seems quite straightforward to assume the conditional independence given the noisy data $\mathbf{x}_t$

$$p(\mathbf{c}_1,...,\mathbf{c}_n|\mathbf{x}_t) = p(\mathbf{c}_1|\mathbf{x}_t)p(\mathbf{c}_2|\mathbf{x}_t)...p(\mathbf{c}_n|\mathbf{x}_t) = \prod_{i=1}^{n} p(\mathbf{c}_i|\mathbf{x}_t),$$

Then, we have the summation formulation to rewrite the gradient term in score function

$$\nabla_{\mathbf{x}_t}\log p(\mathbf{c}_1,...,\mathbf{c}_n|\mathbf{x}_t) = \sum_{i=1}^{n}\nabla_{\mathbf{x}_t}\log p(\mathbf{c}_i|\mathbf{x}_t). \tag{5}$$

Pioneering work [49] uses an energy function to approximate $\log p(\mathbf{c}|\mathbf{x}_t)$ and [4, 47] empirically proposes to use the summation of multiple energy functions to solve the multi-condition guidance problem, which happens to follow the result in Eq. (5).

However, the equation in Eq. (5) can not hold owing to the basic probabilistic theory that independence neither implies conditional independence

$$p(\mathbf{c}_1,...,\mathbf{c}_n) = \prod_{i=1}^{n} p(\mathbf{c}_i) \not\Rightarrow p(\mathbf{c}_1,...,\mathbf{c}_n|\mathbf{x}_t) = \prod_{i=1}^{n} p(\mathbf{c}_i|\mathbf{x}_t).$$

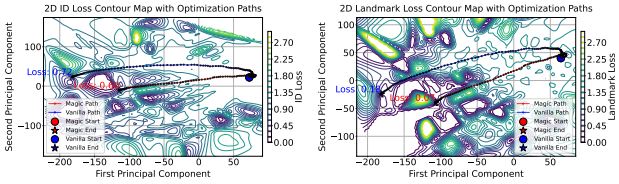
In fact, the noisy data $\mathbf{x}_t$ for $\forall t < n$ contains the information of all conditions involved in the inverse process. For example, if we have already known that $\mathbf{x}_t$ reflects the image of artwork "Sunflowers" polluted by noise, then knowing that $\mathbf{c}_1$ refers to the object of image is sunflowers tells you a lot about the van Gogh's art style, *i.e.*, $\mathbf{c}_2$. Hence, $p(\mathbf{c}_1|\mathbf{x}_t)$ and $p(\mathbf{c}_2|\mathbf{x}_t)$ are not conditionally independent.

3.4. Conditional Dependent Creation

Since the hypothesis on conditional independence is too strong, we consider the real case of conditional dependence where

$$\begin{aligned} p(\mathbf{c}_1,...,\mathbf{c}_n|\mathbf{x}_t) &= p(\mathbf{c}_1|\mathbf{x}_t)p(\mathbf{c}_2|\mathbf{c}_1,\mathbf{x}_t)...p(\mathbf{c}_n|\mathbf{c}_1,...,\mathbf{c}_{n-1},\mathbf{x}_t), \\ &= \prod_{i=1}^{n} p(\mathbf{c}_i|\{\mathbf{c}_{j\in(0,i-1]}\},\mathbf{x}_t). \end{aligned}$$

Sunflowers is the title of two series of still life paintings by the Dutch painter Vincent van Gogh.



(a) ID Loss (b) Landmark Loss

Figure 2. Illustration of the multi-condition optimization landscape with and without the our strategy (best viewed in color). Our approach $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1, \mathbf{c}_2|\mathbf{x}_t)$ navigates the valleys, achieving lower condition losses, whereas the vanilla counterpart $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{x}_t)$ requires more steps to find a decreasing direction, resulting in higher loss values.

To simplify the derivation, we begin with the case with two constraints, i.e., $\{\mathbf{c}_1, \mathbf{c}_2\}$. The controllable term is in the form of

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1, \mathbf{c}_2|\mathbf{x}_t) = \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{x}_t)}_{\text{term I}} + \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{c}_1, \mathbf{x}_t)}_{\text{term II}},$$

where the first term serves as the single condition term as discussed in previous work, and the second term relies on the $\mathbf{c}_1$ guided result, which inspires us first to calculate the intermediate $\hat{\mathbf{x}}_{t,\mathbf{c}_1}$ as

$$\hat{\mathbf{x}}_{t,\mathbf{c}_1} = (2 - \sqrt{1-\beta_t})\mathbf{x}_t + \beta_t \big(\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{x}_t)\big) + \sqrt{\beta_t}\boldsymbol{\epsilon},$$

then use $\hat{\mathbf{x}}_{t,\mathbf{c}_1}$ to derive $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{c}_1, \mathbf{x}_t) \in \mathbb{R}^{1\times HW}$ as

$$\begin{aligned}\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{c}_1, \mathbf{x}_t) &= \nabla_{\mathbf{x}_t} \hat{\mathbf{x}}_{t,\mathbf{c}_1} \cdot \nabla_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}} \log p(\mathbf{c}_2|\hat{\mathbf{x}}_{t,\mathbf{c}_1}), \\ &= \nabla_{\mathbf{x}_t} \hat{\mathbf{x}}_{t,\mathbf{c}_1} \cdot g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}},\end{aligned}$$

where $g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}$ denotes $\nabla_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}} \log p(\mathbf{c}_2|\hat{\mathbf{x}}_{t,\mathbf{c}_1})$ for short and

$$\nabla_{\mathbf{x}_t} \hat{\mathbf{x}}_{t,\mathbf{c}_1} = (2 - \sqrt{1-\beta_t})\mathbf{I} + \beta_t \big(\nabla_{\mathbf{x}_t} \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t) + \mathbf{H}_{\mathbf{x}_t}\big).$$

It is easy to calculate the Jacobian matrix of $\nabla_{\mathbf{x}_t} \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t)$ via standard deep learning libraries such as TensorFlow[1]/PyTorch[33]. However, the Jacobian and Hessian matrix $\mathbf{H}_{\mathbf{x}_t} = \nabla^2_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{x}_t)$ can be the bottleneck of computation cost and memory footprint . Fortunately, by utilizing the product of $\nabla_{\mathbf{x}_t} \hat{\mathbf{x}}_{t,\mathbf{c}_1}$ and $g^T_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}$, we can avoid the direct calculation of the Hessian matrix.

To simplify the derivation, we denote $(2 - \sqrt{1-\beta_t})\mathbf{I} + \beta_t \nabla_{\mathbf{x}_t} \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t)$ as matrix $\mathbf{A}$, then

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{c}_1, \mathbf{x}_t) = \mathbf{A} \cdot g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}} + \beta_t \mathbf{H}_{\mathbf{x}_t} \cdot g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}. \quad (6)$$

Taking image with 256×256 resolution as an example, the Hessian matrix consumes $512^4 \times 4$ bytes, i.e., 256GB, which is impractical.

Similar to $g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}$, we further rewrite $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{x}_t)$ as $g_{\mathbf{x}_t}$, then we have

$$\begin{aligned}\mathbf{H}_{\mathbf{x}_t} \cdot g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}} = \frac{\partial g^T_{\mathbf{x}_t}}{\partial \mathbf{x}_t} \cdot g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}} &= \frac{\partial g^T_{\mathbf{x}_t}}{\partial \mathbf{x}_t} \cdot g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}} + g^T_{\mathbf{x}_t} \cdot \frac{\partial g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}}{\partial \mathbf{x}_t} \\ &= \frac{\partial (g^T_{\mathbf{x}_t} g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}})}{\partial \mathbf{x}_t}. \quad (7)\end{aligned}$$

Here, we make use of the trick that $\frac{\partial g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}}{\partial \mathbf{x}_t} = \mathbf{0}$, which convert the product between the Hessian matrix and gradient vector to the gradient of $\mathbf{x}_t$ with respect to a scalar $g^T_{\mathbf{x}_t} g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}$. The computing cost and GPU memory footprint can be much lower than the vanilla formulation. The conclusion also holds for $\nabla_{\mathbf{x}_t}\big(\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t) \cdot g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}\big)$ in $\mathbf{A} \cdot g_{\hat{\mathbf{x}}_{t,\mathbf{c}_1}}$.

We summarize the pipeline with PC sampler [42] to solve Equation (1) in Algorithm 1. Note that our approach only modifies the sampling process to apply the multiple attribute generation leaving out the training procedure, which naturally fits the need of zero-shot learning. Figure 2 further illustrates the superiority of Algorithm 1 compared to conditional independent generation. We randomly sample ten thousands results from the space of $\mathbf{x}_t$, then project them to 2D via PCA to construct the loss landscape.

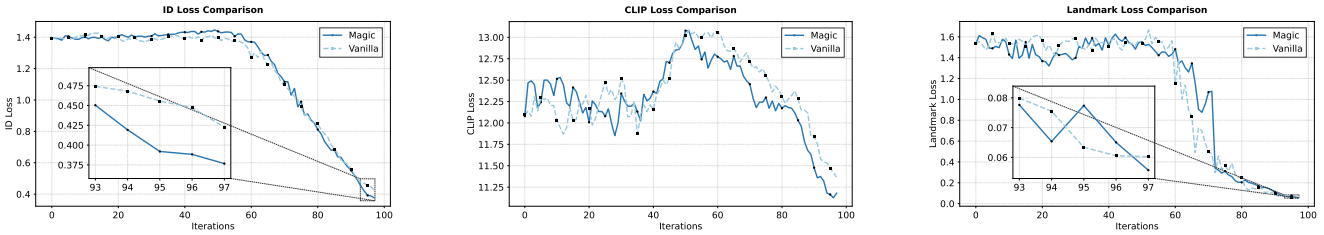
Algorithm 1 Conditional Sampling

$\mathbf{x}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = N$ **to** 1 **do**
  $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{c}_1) \leftarrow \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{x}_t)$
  $\hat{\mathbf{x}}_{t,\mathbf{c}_1} \leftarrow (2 - \sqrt{1-\beta_t})\mathbf{x}_t + \beta_t \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{c}_1)$
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\hat{\mathbf{x}}_{t,\mathbf{c}_1} \leftarrow \hat{\mathbf{x}}_{t,\mathbf{c}_1} + \sqrt{\beta_t}\mathbf{z}$
  $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{c}_1, \mathbf{x}_t) \leftarrow$ solve Eq.(6) with Eq.(7)
  $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1, \mathbf{c}_2|\mathbf{x}_t) \leftarrow \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{c}_1, \mathbf{x}_t)$
  $\mathbf{x}_{t-1} \leftarrow (2 - \sqrt{1-\beta_t})\mathbf{x}_t + \beta_t \big(\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1, \mathbf{c}_2|\mathbf{x}_t)\big)$
  $\mathbf{x}_{t-1} \leftarrow \mathbf{x}_{t-1} + \sqrt{\beta_t}\mathbf{z}$
return $\mathbf{x}_0$

3.5. A Multi-task Learning Perspective

Though Section 3.4 alleviates the image customization with two conditions, the complexity for solving $p(\mathbf{c}_i|\mathbf{c}_1, ..., \mathbf{c}_{i-1}, \mathbf{x}_t)$ still increases with the number of conditions. For example, $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_3|\mathbf{c}_2, \mathbf{c}_1, \mathbf{x}_t)$ should introduce a third-order derivative, which corresponds to a 3D tensor. Therefore, we propose a simple heuristic method to solve the multi-condition customization.

Given multiple conditions $\{\mathbf{c}_1, ..., \mathbf{c}_n\}$, we may go through every $\mathbf{c}_i$ from $\mathbf{c}_1$ to $\mathbf{c}_n$ as the initial condition. Then, follow the derivation in Section 3.4, we obtain $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_j, \mathbf{c}_i|\mathbf{x}_t)$ for $\forall i, j$. If we regard $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1, \mathbf{c}_2|\mathbf{x}_t)$ as the gradient descent step for the



(a) ID Loss Variation Across Iterations (b) CLIP Loss Variation Across Iterations (c) Landmark Loss Variation Across Iterations

Figure 3. We visualize the loss curves for generating a face image under three guided conditions using $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3|\mathbf{x}_t)$, comparing results with vanilla $\sum_i \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_i|\mathbf{x}_t)$. The proposed method achieves relatively lower metric values across all conditions, indicating a better balance in multi-condition control.

Algorithm 2 Multi-condition Approximation

$c \in [0, 1)$
for $i = 1$ **to** n **do**
  for $j = 1$ **to** n **do**
    $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_j|\mathbf{c}_i, \mathbf{x}_t) \leftarrow$ solve Eq.(6) with Eq.(7)
    $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_i, \mathbf{c}_j|\mathbf{x}_t) \leftarrow \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_i|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_j|\mathbf{c}_i, \mathbf{x}_t)$
$g_0 \leftarrow \frac{1}{n(n-1)} \sum_i \sum_j \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_i, \mathbf{c}_j|\mathbf{x}_t)$
$\phi \leftarrow c^2 ||g_0||^2$
$g_w \leftarrow \frac{1}{n(n-1)} \sum_i \sum_j w_{i,j} p(\mathbf{c}_i, \mathbf{c}_j|\mathbf{x}_t)$
$w^* = \arg\min_w g_w^T g_0 + \sqrt{\phi} ||g_w||$
return $g_0 + \frac{\sqrt{\phi}}{||g_{w^*}||} g_{w^*}$

loss function $\min -\log p(\mathbf{c}_1, \mathbf{c}_2|\mathbf{x}_t)$, then the optimal gradient step $g_{\mathbf{x}_t}$ satisfying all losses $\min -\log p(\mathbf{c}_j, \mathbf{c}_i|\mathbf{x}_t)$ where $\{\forall i, j \in [1, n]\}$ serves as an approximation to $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1, ..., \mathbf{c}_n|\mathbf{x}_t)$. Formally, the optimal gradient step $g_{\mathbf{x}_t}$ solves the following goal

$$\mathbf{x}_t^* = \arg\min \left\{ \mathcal{L}(\mathbf{x}_t) \triangleq \sum_i \sum_j -\log p(\mathbf{c}_j, \mathbf{c}_i|\mathbf{x}_t) \right\}, \quad (8)$$

which is the standard formulation of Multi-Task Learning (MTL) [43]. It uses $g_{\mathbf{x}_t}$ to find an optimal $\mathbf{x}_t^*$ that achieves low losses across all tasks.

In this paper, we adopt conflict-averse gradient descent (CAGrad [24]) to tackle the MTL problem. Algorithm 2 further illustrates how to approximate $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1, \ldots, \mathbf{c}_n|\mathbf{x}_t)$ by CAGrad. Consequently, Figure 3 demonstrates that our approach performs effectively in multi-condition generation.

3.6. Choice of Condition Classifier $p(\boldsymbol{c}|\boldsymbol{x}_t)$

The remaining challenge in implementing Algorithms 1 and 2 lies in determining $p(\mathbf{c}|\mathbf{x}_t)$. A common approach for class-conditional sampling involves employing a pre-trained, time-dependent classifier model for $p(\mathbf{c}|\mathbf{x}_t)$ [13, 31, 42], or using a handcrafted energy function for conditional control [26, 49], where $\nabla \mathcal{E}(\mathbf{c}, \mathbf{x}, t) \propto \nabla \log p(\mathbf{c}|\mathbf{x}_t)$. Recent studies [4, 6, 28, 32, 47] leverage the predicted clean image $\mathbf{x}_{0|t}$ to eliminate the dependency on timestamp t, using $\nabla \mathcal{E}(\mathbf{c}, \mathbf{x}_{0|t})$ as an approximation of $\nabla \log p(\mathbf{c}|\mathbf{x}_t)$. While our approach is compatible with all these strategies, the latter is particularly advantageous for our multi-attribute customization, as it functions as a plug-and-play module that circumvents the need to explicitly train $p_t(\mathbf{c}_i|x_t)$ with $x_t \sim p_t(\mathbf{x}_t|\mathbf{c}_j)$. Consequently, we focus on using a differentiable guidance function f, such as a pre-trained perceptual model for image classification, detection, or segmentation, to represent $\mathcal{E}(\mathbf{c}, \mathbf{x}_{0|t})$.

4. Experiments

4.1. Implementation Details

We apply our approach to various pre-trained open-source diffusion models to enable zero-shot manipulation of multiple attributes. Following the settings of recent conditional creation methods [6], we perform ancestral sampling as proposed in [18] with 100 steps. We focus on multi-modal face generation using an unconditional human face diffusion model [29] as the base model and stylized, text-guided image creation based on the stable diffusion model [38] to ensure fair comparisons with other approaches. Attribute control is achieved by integrating text, segmentation, landmarks, face ID, content, and style information. To address potential challenges with learning rate selection in score-based conditional models, we adopt the same settings reported in [47] to avoid extensive search. As for the formulation of condition classifier, we use the following strategy:
Text: We employ the pre-trained CLIP model [35], which includes both text and image encoders (with the input size of 224×224), to project the generated image and guide-text into a shared feature space. To align these representations, we minimize the cosine similarity between them.
Segmentation: For segmentation, we utilize the face parsing network [51] to generate segmentation maps for both the input reference image and the denoised image. We then apply Mean Squared Error (MSE) loss to minimize the distance between these segmentation maps.

a) ID + Landmark (1st. Row) *v.s.* Landmark + ID (2nd. Row)

b) Seg. map + Text (3rd. Row) *v.s.* Text + Seg. map (4th. Row)

Figure 4. Illustration of the effect of different condition sequences. When two conditions modify the same facial features, as shown in a), the sequence becomes important. However, if the conditions are weakly correlated, such as hair color and face parsing, the sequence has minimal impact as shown in b).

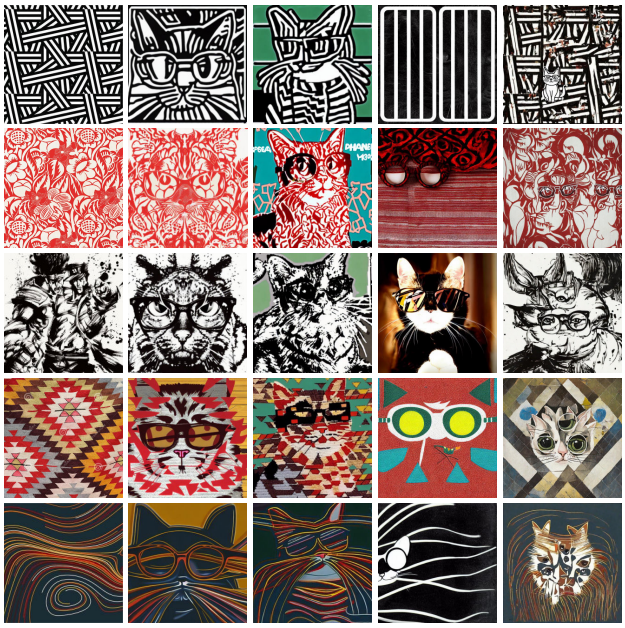


Figure 5. Visualization of image creation guided by text and style reference images. All images are generated using the prompt "cat wearing glasses," with style reference images displayed in the first column.

Landmark: Using a well-known open-source framework for human face landmark detection [7], we identify precise landmark positions based on face detection results. As with segmentation, we compute the Euclidean distance between the landmarks on the target and generated images.

Face ID: To capture discriminative facial features, we leverage ArcFace [8]. In line with common practice for facial similarity assessments, we calculate Cosine similarity to determine the match between two faces.

Content: For content creation, we use the standard v1.4 stable diffusion model [38] as our base and its output size is 512×512. This model generates images guided by a prompt, or if no specific guidance is required, we use a wildcard ("*") to produce an unconditioned output.

Style: To ensure stylistic consistency, we draw on the neural style transfer framework [14]. Here, the style similarity is quantified by comparing the Gram matrices produced by the image encoder for both the reference and generated images.

4.2. Effect of Condition Sequence

We conduct an ablation study on the impact of using different condition sequences, such as $\{\mathbf{c}_1, \mathbf{c}_2\}$ or $\{\mathbf{c}_2, \mathbf{c}_1\}$, during the diffusion process. Specifically, we examine the differences between applying $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{c}_2, \mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{x}_t)$ versus $\nabla_{\mathbf{x}_t} \log p(\mathbf{c}_2|\mathbf{c}_1, \mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{c}_1|\mathbf{x}_t)$ to the denoised intermediate result. As shown in Figure 4, the results reveal that if both conditions strongly influence the same facial features, we adhere to the preference of the pre-trained model. For example, when applying face ID and landmark conditions, we find that establishing face ID first for a facial generation model pre-trained on CelebA and then refining landmarks contributes to superior outcomes. If the attributes do not directly conflict, the sequence has minimal effect.

4.3. Evaluation

In this section, we present both quantitative and qualitative evaluations. We assess the generated results using qualitative metrics, including Frechet Inception Distance (FID) [16] calculated on CelebA-HQ [23], as well as the distance to the given conditions (ID, segmentation, landmarks, and text) using five hundreds randomly selected samples coupled with corresponding reference images, labels and prompts.

Face Segmentation, ID and Text Control: As with dual-condition generation, the independent assumption used in [47] becomes even less realistic in three-condition (text, face parsing and face ID) guided creation. Consequently, our approach (i.e., Algorithm 2) outperforms the baseline method, as shown in Figure 6 and Table 1. The text distance appears inconsistent with Figure 3 because we did not apply condition control in the final steps, resulting in a slightly larger final text distance.

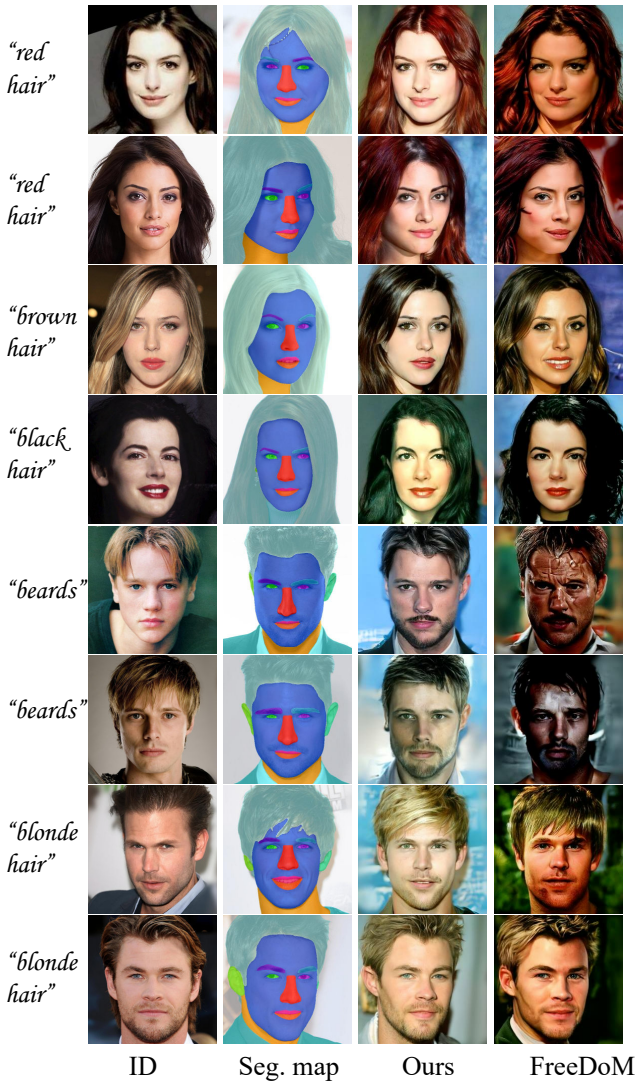


Figure 6. Visualization of face synthesis guided by text, face ID and parsing conditions.

Method	FID ↓	Seg. Dist. ↓	ID Dist. ↓	Text Dist. ↓
FreeDoM [47]	136	1771	0.501	0.774
Ours	**123**	**1677**	**0.475**	**0.769**

Table 1. Quantitative comparison with a condition-independent baseline, evaluating FID, segmentation (Seg.) distance, identity (ID) distance, and text distance. Lower values indicate better performance.

Text and Style Control: When coupled with the stable diffusion model, we treat prompt-guided generation as $p(\mathbf{x}_t|\mathbf{c}_1)$ and then apply the style condition $\mathbf{c}_2$ to complete the derivation. As shown in Figures 5 and 7, our approach continues to perform well even under highly abstract style conditions. For comparison, we use two popular methods, UGD [4] and StyleAligned [15], as baseline methods. Our approach captures the patterns and color schemes of the reference images more effectively, reflecting greater creativity. Quantitative results are shown in Table 2.

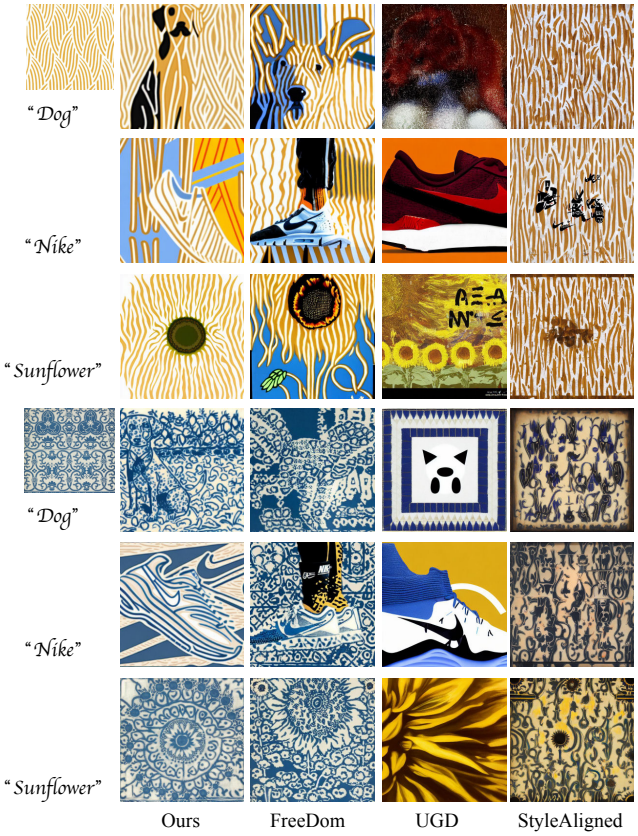


Figure 7. Visualization of image creation guided by text and style reference images. We use various prompts and style reference images, shown in the first column, to perform conditional image generation.

Method	Content Loss ↓	Style Loss ↓	Text Distance ↓
StyleAligned [15]	-	11.35	0.7475
UGD [4]	-	18.04	0.7682
FreeDoM [47]	1.93	10.21	0.7156
Ours	**1.82**	**10.14**	**0.7152**

Table 2. Quantitative comparison with other approaches based on stylized results. We use plain SD-v1.4 results with prompts but without style priors as original content images to calculate content loss. Style loss follows the standard settings in style transfer methods, and text loss is computed using the CLIP model.

Face Segmentation and Text Control: We compare our method with state-of-the-art approaches, including Collaborative Diffusion [19], TediGAN [46], and FreeDoM [47], using segmentation and text guidance as conditions, as shown in Figure 8 and Table 3. Although TediGAN achieves a high FID score, it frequently generates face images that do not match the segmentation mask or deviate

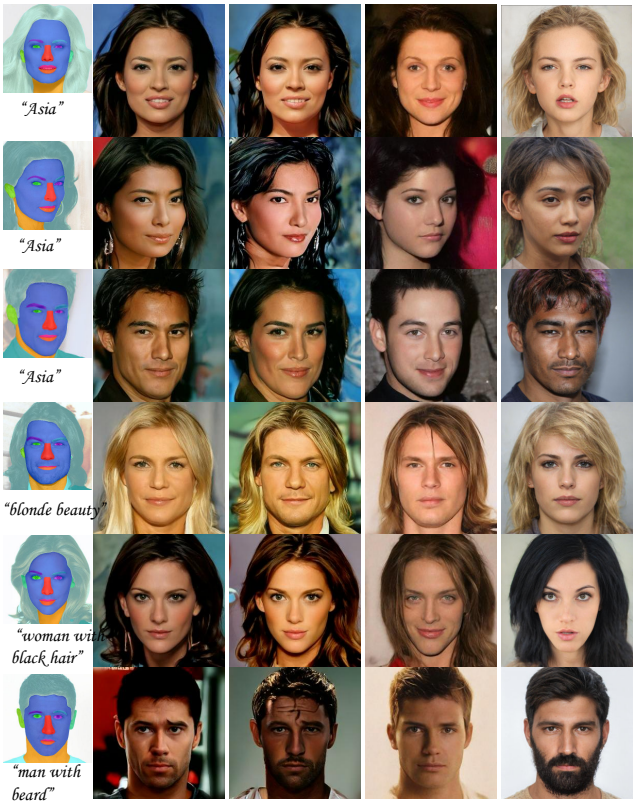


Figure 8. Visualization of face synthesis guided by text and face parsing conditions.

Method	FID ↓	Text Distance ↓	Seg. Distance ↓
TediGAN [46]	**98**	0.763	2344
Collaborative [19]	122	0.758	2162
FreeDoM [47]	136	0.764	1569
Ours	127	**0.752**	**1502**

Table 3. Quantitative comparison of recent open-source facial conditional generation methods, assessing Fréchet Inception Distance (FID), text embedding distance, and segmentation distance.

from the specified guide text. Collaborative Diffusion performs slightly better but still fails to consistently follow the instructions in certain cases.

Face ID and Landmark Control: Figure 9 presents our visual results controlled by Face ID and landmark positions, compared with the conditional independent baseline method, FreeDoM. While FreeDoM can generate face images that align with the input conditions, it often struggles to balance multiple attributes effectively as shown in Table 4. The SOTA face swapping methods [27, 50] also result in issues such as weakened ID consistency or inaccurate landmark adherence, though, landmark is required during generation. [50] fails to consider the target face, resulting in a

Figure 9. Visualization of face synthesis guided by face ID and landmark conditions.

Method	FID ↓	Landmark Distance ↓	ID Distance ↓
DiffSwap [50]	119	**0.103**	1.167
E4S [27]	**92**	0.282	0.977
FreeDoM [47]	134	0.195	0.740
Ours	124	0.194	**0.549**

Table 4. Quantitative results for the two tasks, ID + Landmark, evaluating FID, landmark distance, and identity (ID) distance. Lower values indicate better performance.

high landmark score, since the original face without ID conditioning should yield a zero landmark distance. Although [27] achieves high-fidelity results, it does so at the cost of partially disregarding both conditions.

5. Conclusion

This paper introduced a novel approach for multi-attribute guided generation that explicitly models the conditional dependencies among attributes, enhancing coherence in generated images. Unlike traditional models that assume conditional independence, our method uses a score-based conditional diffusion model to sequentially condition attributes, refining each with contextually-aware gradients. Experimental results demonstrate that Z-Magic effectively overcomes inconsistencies inherent in independent conditioning, achieving high coherence and computational efficiency.

Acknowledgements. This work was partly supported by Beijing Science and Technology Plan Project under No. Z231100005923033 and L221013.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*, pages 265–283. USENIX Association, 2016. 4

[2] Brian. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12: 313–326, 1982. 3

[3] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18370–18380. IEEE, 2023. 1

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 3, 5, 7

[5] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14347–14356. IEEE, 2021. 1

[6] Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 5

[7] cunjian. pytorch_face_landmark, 2023. Accessed: 2024-11-06. 6

[8] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 6

[9] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *ACM International Conference on Multimedia*, pages 2719–2727, 2020. 2

[10] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1210–1217, 2021.

[11] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11326–11336, 2022.

[12] Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. Z*: Zero-shot style transfer via attention reweighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 6934–6944. IEEE, 2024. 2

[13] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794, 2021. 2, 5

[14] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. 6

[15] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 4775–4785. IEEE, 2024. 1, 2, 7

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 6

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3, 5

[19] Ziqi Huang, Kelvin C. K. Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6080–6090. IEEE, 2023. 1, 2, 7, 8

[20] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005. 3

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021. 2

[22] Jihyun Kim, Changjae Oh, Hoseok Do, Soohyun Kim, and Kwanghoon Sohn. Diffusion-driven GAN inversion for multi-modal face image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 10403–10412. IEEE, 2024. 1, 2

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *2020 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 5548–5557. Computer Vision Foundation / IEEE, 2020. 6

[24] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18878–18890, 2021. 5

[25] Renshuai Liu, Bowen Ma, Wei Zhang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, and Xuan Cheng. Towards a simultaneous and granular identity-expression control in personalized face generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 2114–2123. IEEE, 2024. 2

[26] 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 *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 289–299. IEEE, 2023. 5

[27] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. Fine-grained face swapping via regional GAN inversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8578–8587. IEEE, 2023. 8

[28] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11451–11461. IEEE, 2022. 5

[29] 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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 5

[30] 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 *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 4296–4304. AAAI Press, 2024. 2

[31] 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 *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 16784–16804. PMLR, 2022. 2, 5

[32] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, pages 11:1–11:11. ACM, 2023. 5

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. 4

[34] 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. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 1

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 2, 5

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8821–8831. PMLR, 2021. 1

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 1, 2, 5, 6

[39] Jonathan Richard Shewchuk. High dimensions; random projection; the pseudoinverse, 1998. https://people.eecs.berkeley.edu/~jrs/189/lec/22.pdf. 2

[40] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, pages 574–584. AUAI Press, 2019. 3

[41] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing*

Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pages 1415–1428, 2021.

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3, 4, 5

[43] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(7):3614–3633, 2022. 5

[44] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 6232–6242. IEEE, 2024. 1

[45] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18051–18060. IEEE, 2022. 2

[46] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2256–2265. Computer Vision Foundation / IEEE, 2021. 2, 7, 8

[47] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 23117–23127. IEEE, 2023. 3, 5, 6, 7, 8

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3813–3824. IEEE, 2023. 1, 2

[49] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. EGSDE: unpaired image-to-image translation via energy-guided stochastic differential equations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 3, 5

[50] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8568–8577. IEEE, 2023. 8

[51] zllrunning. face-parsing.pytorch, 2023. Accessed: 2024-11-06. 5