Conditional Image Synthesis with Diffusion Models: A Survey

Anonymous authors
Paper under double-blind review

Abstract

Conditional image synthesis based on user-specified requirements is a key component in creating complex visual content. In recent years, diffusion-based generative modeling has become a highly effective way for conditional image synthesis, leading to exponential growth in the literature. However, the complexity of diffusion-based modeling, the wide range of image synthesis tasks, and the diversity of conditioning mechanisms present significant challenges for researchers to keep up with rapid developments and understand the core concepts on this topic. In this survey, we categorize existing works based on how conditions are integrated into the two fundamental components of diffusion-based modeling, *i.e.*, the denoising network and the sampling process. We specifically highlight the underlying principles, advantages, and potential challenges of various conditioning approaches in the training, re-purposing, and specialization stages to construct a desired denoising network. We also summarize six mainstream conditioning mechanisms in the essential sampling process. All discussions are centered around popular applications. Finally, we pinpoint some critical yet still open problems to be solved in the future and suggest some possible solutions.

1 Introduction

Image synthesis is an essential generative AI task. It is more useful when incorporating user-provided conditions to generate images that meet diverse user needs through precise control. Early works have made significant breakthroughs in various conditional image synthesis tasks, such as text-to-image generation Reed et al. (2016); Zhang et al. (2017); Ding et al. (2021); Ramesh et al. (2021), image restoration Ledig et al. (2017); Wang et al. (2021); Maaløe et al. (2019); Lee et al. (2022), and image editing Brock et al. (2017); Ling et al. (2021); Abdal et al. (2020). However, the performance of conditional image synthesis with early deep learning-based generative models such as generative adversarial networks (GANs) Goodfellow et al. (2014); Mirza & Osindero (2014), variational auto-encoders (VAEs) Kingma & Welling (2014); Sohn et al. (2015), and auto-regressive models (ARMs) Van Den Oord et al. (2016); Van den Oord et al. (2016) is unsatisfactory due to their intrinsic limitations: GANs are vulnerable to mode collapse and unstable training Goodfellow et al. (2014); VAEs often generate blurry images Kingma & Welling (2014); and ARMs suffer from sequential error accumulation and huge time consumption Van Den Oord et al. (2016).

In recent years, diffusion models (DMs) have emerged as state-of-the-art image generation models due to their strong generative capabilities and versatility Sohl-Dickstein et al. (2015); Ho et al. (2020); Song et al. (2021b); Karras et al. (2022); Chen et al. (2024a). In DMs, images are synthesized from Gaussian noise through iterative denoising steps guided by the predictions of a denoising network. This distinctive multi-step sampling process enables DMs to achieve remarkable generative performance characterized by stable training, diverse outputs, and exceptional sample quality. It also gives DMs a unique advantage in facilitating conditional integration compared to one-step generative models. These benefits have made DMs the tool of choice for conditional image synthesis, leading to rapid growth in the research on *Diffusion-based Conditional Image Synthesis* (DCIS) over the past few years Rombach et al. (2022); Saharia et al. (2022b); Lu et al. (2023); Choi et al. (2021); Saharia et al. (2022c); Kawar et al. (2023); Hertz et al. (2023); Zhang et al. (2023e); Gal et al. (2023a); Zhang et al. (2023b); Wang et al. (2024a). Fig. 1 illustrates seven popular DCIS tasks with various input modalities.

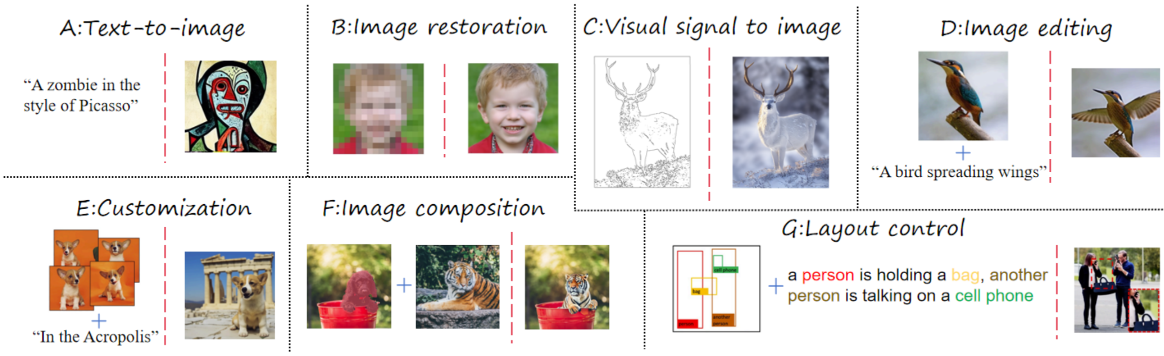

Figure 1: Seven representative conditional image synthesis tasks with their input/output. Figures are cited from the following papers: (A) Stable Diffusion Rombach et al. (2022); (B) SR3 Saharia et al. (2022c); (C) ControlNet Zhang et al. (2023b); (D) Imagic Kawar et al. (2023); (E) DreamBooth Ruiz et al. (2023); (F) PbE Yang et al. (2023a); (G) InteractDiffusion Hoe et al. (2023).

The rapidly expanding body of works, the numerous variations in model architectures, training methods, and sampling techniques, along with the broad scope of potential conditional synthesis tasks, make it challenging for researchers to grasp the full landscape of DCIS. This complexity can be particularly overwhelming for newcomers to the field. What is needed is a systematic survey that offers a comprehensive yet structured overview of this growing research area.

There exist several surveys on specific conditional image synthesis tasks, such as image restoration Li et al. (2023g), text-to-image Zhang et al. (2023a), and image editing Huang et al. (2024), or classifying works in computer vision according to their target conditional synthesis tasks Croitoru et al. (2023); Po et al. (2023). While these task-oriented surveys provide valuable insights into approaches for their respective target tasks, they do not include the commonalities in model frameworks across different conditional synthesis tasks in terms of model architectures and conditioning mechanisms. Two recent surveys Shuai et al. (2024); Cao et al. (2024) provide overview on DM-based works for a wide range of tasks in the field of conditional image synthesis. However, their scope remains limited as they primarily focus on DCIS works built on T2I backbones, neglecting earlier works that integrate conditioning into unconditional denoising networks or involve training task-specific conditional denoising networks from scratch. These earlier efforts are foundational for the current advancements in DCIS using T2I backbones and are still widely applied in low-level tasks such as image restoration. Besides, Shuai et al. (2024) focuses most of its attention on the DM-based image editing framework and lacks systematic analysis on the unified framework for other tasks in this field while Cao et al. (2024) does not delve deeper into the design choices in model architecture and detailed conditioning mechanisms for sampling process. This leads to a lack of systematization in their taxonomies and the omission of crucial related works in the field of DCIS.

In contrast, this survey aims to provide a comprehensive and structured framework that covers a wide range of current DCIS works by offering a taxonomy based on the mainstream techniques for condition integration in DCIS frameworks. We present a clear and systematic breakdown of the components and design choices involved in constructing a DCIS framework with condition integration. Specifically, we review and summarize existing DCIS methods by examining how conditions are integrated into the two fundamental components of diffusion modeling: the *denoising network* and the *sampling process*. For the denoising network, we break down the process of establishing a conditional denoising network into three stages. For the sampling process, we categorize six mainstream in-sampling conditioning mechanisms, detailing how control signals are integrated into various components of the sampling process. The objective is to give readers a high-level and accessible overview of existing DCIS works across diverse tasks, equipping them to design conditional synthesis frameworks for their own desired tasks, including novel tasks that have yet to be explored.

The remainder of this survey is organized as follows: we first introduce the background of diffusion models and the conditional image synthesis task in Sec. 2. Next, we summarize methods for condition integration within the denoising network in Sec. 3, and for the sampling process in Sec. 4. Finally, we explore potential future directions in Sec. 5.

2 Backgrounds

Diffusion-based generative modeling adopts a forward diffusion process of gradually adding noise into clean data and learns a denoising network to predict the added noise. In the sampling process, data is synthesized by reversing the forward process from Gaussian noise based on the prediction of a denoising network. We first introduce the core concepts of discrete-time and continuous-time diffusion modeling in Sec. 2.1. Then, we discuss the model architecture in Sec. 2.2 and highlight representative DCIS tasks in Sec. 2.3.

2.1 The Formulation of Diffusion Modeling

2.1.1 Discrete-Time Formulation

The discrete-time diffusion model was initially proposed in Sohl-Dickstein et al. (2015). It constructs a forward Markov chain to transform clean data into noise by progressively adding small amounts of Gaussian noise so that a parameterized denoising network can be learned to predict the added noise in each forward step. Once the denoising network is trained, images can be generated from Gaussian noise by reversing the diffusion process. This idea gained popularity through an important follow-up work known as denoising diffusion probabilistic models (DDPMs) Ho et al. (2020). This work led to a substantial improvement in the quality of synthesized images and increased resolutions, from 32×32 Sohl-Dickstein et al. (2015) to 256×256, sparking a rapidly growing interest in diffusion models. Next, we adopt the notation from DDPM Ho et al. (2020), which is widely used in the literature to describe discrete-time diffusion models Song et al. (2021a); Rombach et al. (2022); Kawar et al. (2023).

The forward Markov chain is parameterized based on a pre-defined schedule $\beta_1, \ldots, \beta_T$, where β_t is the noise variance in each step and the total number of steps T is usually large, *e.g.*, 1,000. Given the clean data sampled from the training dataset $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$, the transition kernel is $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right)$, or, $q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t) I\right)$, where $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are latent variables, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, and $\bar{\alpha}_T \to 0$. By progressively adding Gaussian noise to the clean data, this Markov chain transforms the data distribution to an approximate normal distribution, *i.e.*, $\int q(\mathbf{x}_T|\mathbf{x}_0) p_{\text{data}}(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \approx \mathcal{N}(0, \mathbf{I})$.

In the training phase, DDPM Ho et al. (2020) learns a denoising network with parameter $\boldsymbol{\theta}$ by minimizing the KL divergence between the transition kernel $p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and the posterior distribution $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$. In practice, DDPM Ho et al. (2020) is trained on the following re-parameterized loss function to improve the training stability and sample quality:

$$\mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right], \tag{1}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ is a noise-prediction network to estimate the added noise $\boldsymbol{\epsilon} = \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{\sqrt{1-\bar{\alpha}_t}}$ in each step. For the conditional generation that performs denoising steps conditioned on control signal $\mathbf{c}$, the conditional denoising network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c})$ can be trained on a loss function similar to Eq. 1.

In the sampling process, DDPM gradually generates clean data from Gaussian noise by computing the reverse transition kernel $p_{\boldsymbol{\theta}}$ with the learned network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$, i.e.,

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\right) + \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \boldsymbol{\epsilon}_t, \tag{2}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{I})$ is the standard Gaussian noise independent of $\mathbf{x}_t$. The following work DDIM Song et al. (2021a) proposed a family of sampling processes sharing the same marginal distribution $p(\mathbf{x}_t)$ with the above sampling process, which are written as

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_t) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \cdot \boldsymbol{\epsilon}_{\boldsymbol{\theta}} + \sigma_t \boldsymbol{\epsilon}_t, \tag{3}$$

where $\boldsymbol{f_\theta}\left(\mathbf{x}_t\right) = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon_\theta}}{\sqrt{\bar{\alpha}_t}}$ denotes the predicted $\mathbf{x}_0$ at time step t. For simplicity, we will refer to $\boldsymbol{f_\theta}(\mathbf{x}_t)$ as the intermediate denoising output $\mathbf{x}_{0|t}$ hereafter. Each choice of σ_t represents a specific sampling process in DDIM Song et al. (2021a). It is identical to the DDPM generative process in Eq. 2 when $\sigma_t = \sqrt{\left(1-\bar{\alpha}_{t-1}\right)/\left(1-\bar{\alpha}_t\right)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}$ and becomes a deterministic process when $\sigma_t = 0$.

2.1.2 Continuous-Time Formulation

Song et al. Song et al. (2021b) proposed to formulate a diffusion process $\{\mathbf{x}_t \sim p_t(\mathbf{x})\}_{t=0}^{T}$ with the continuous time variable $t \in [0, T]$ as the solution of an Itô stochastic differential equation (SDE) $d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}_t$, where $\mathbf{w}_t$ denotes the standard Wiener process, and $\mathbf{f}(\mathbf{x}, t)$ and $g(t)$ are drift and diffusion coefficients, respectively Oksendal (2013); Chen et al. (2023). This diffusion process smoothly transforms a data distribution into an approximate noise distribution p_n and its specific discretization recovers the forward process of DDPM Ho et al. (2020). There exists a probability flow ordinary differential equation (PF-ODE) $\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g(t)^2\nabla_{\mathbf{x}}\log p_t(\mathbf{x})\right]\mathrm{d}t$, sharing the same marginal distribution with the reverse SDE $d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2\nabla_{\mathbf{x}}\log p_t(\mathbf{x})\right]dt + g(t)d\hat{\mathbf{w}}$ Song et al. (2021b); Karras et al. (2022); Zhang & Chen (2023); Chen et al. (2024a). Therefore, we can learn a time-dependent score-based denoising network $\mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_t, t\right)$ to estimate the score function $\nabla_{\mathbf{x}_t}\log p\left(\mathbf{x}_t\right)$ with a sum of denoising score matching Vincent (2011); Lyu (2009) objectives weighted by $\lambda(t)$:

$$\mathbb{E}_t\left[\lambda(t)\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_t, t\right) - \nabla_{\mathbf{x}}\log p\left(\mathbf{x}_t \mid \mathbf{x}_0\right)\right\|_2^2\right]\right]. \tag{4}$$

When the score-based denoising network $\mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_t, t\right)$ is trained, we can employ general-purpose numerical methods such as Euler-Maruyama and Runge-Kutta methods to solve the reverse SDE or PF-ODE and recover clean data $\mathbf{x}_0$ from $\mathbf{x}_T$.

In the following sections, unless otherwise specified, we will use notation $\boldsymbol{\epsilon_\theta}$ to represent the denoising network.

2.2 Architecture of the Denoising Network

Pioneering works adopted U-Net Ronneberger et al. (2015) as the denoising network architecture Ho et al. (2020); Song et al. (2021a); Song & Ermon (2019; 2020). A U-Net typically consists of an U-shaped encoder-decoder structure with skip connections. The encoder leverages a stack of residual layers and downsampling convolutions to reduce the spatial data dimension and the decoder upsamples the compressed data back to the original dimension. The U-Net architecture is advantageous for diffusion models due to its exceptional feature extraction, contextual understanding, precise segmentation, and dimensionality preservation property, which enables accurate modeling of complex data distributions for high-quality synthesis. Many followed-up works improved the vanilla U-Net architecture by incorporating multi-head attention Song et al. (2021b); Dhariwal & Nichol (2021); Nichol & Dhariwal (2021), normalization Ho et al. (2020); Dhariwal & Nichol (2021); Nichol & Dhariwal (2021), or cross-attention layersRombach et al. (2022); Saharia et al. (2022b). Recently, transformers emerged as an alternative for denoising networks because of its capability in capturing long-range dependencies Dosovitskiy et al. (2020); Li et al. (2023d), and have achieved success in many tasks including class-conditional generation Yang et al. (2022b), text-to-image generation Bao et al. (2023); Sheynin et al. (2023); Tang et al. (2022); Gu et al. (2022); Li et al. (2023d), layout generation Chai et al. (2023), and medical image generation Pan et al. (2023a).

In the following sections, unless otherwise specified, we assume the architecture of the denoising network adopts a U-Net structure.

2.3 Conditional Image Synthesis Tasks

A conditional image synthesis task $\mathcal{T}$ generates target image $\mathbf{x}$ by sampling from a conditional distribution:

$$\mathbf{x} \sim p_{\mathcal{T}}(\mathbf{x}|\mathbf{c}),\ \mathbf{c} \in \mathcal{D}_{\mathcal{T}}, \tag{5}$$

where $\mathcal{D}_{\mathcal{T}}$ is the domain of conditional input $\mathbf{c}$, and $p_{\mathcal{T}}$ is the conditional distribution defined by the task $\mathcal{T}$. Based on the form of conditional inputs and the correlation between the conditional input and the target

Table 1: Stack of conditioning mechanisms of mainstream synthesis tasks applied to denoising network and sampling process, respectively. Conditioning encoder indicates the module to convert conditional inputs into task-related feature embedding, where * indicates that the encoder is determined by the specific restoration task. ♠, ♡, ♣, ♢ denote the four re-purposing stage condition injection methods described in Sec. 3.2.2. Due to page width limitations, we have placed the DCIS works performing condition integration via the presented stacks of conditioning mechanisms in the row identified by corresponding serial numbers in Tab. 2 in appendix.

Stack of conditioning mechanisms for denoising network						
Task	**Training (backbone)**	**Conditional encoder**	**Condition Injection**	**Backbone fine-tuning**	**Specialization**	**Serial Number**
Text-to-image	✓	CLIP, BERT, LLMs	♡	✗	✗	DN1
Image restoration	✓	Non.	♠	✗	✗	DN2
	✓	*	♠, ♡	✗	✗	DN3
Image editing	✗ (T2I DM)	LLMs-based	♡	✓	✗	DN4
	✗ (T2I DM)	Non.	♠	✓	✗	DN5
	✗ (T2I DM)	Non./BLIP	♡	✗	✓	DN6
Customization	✗ (T2I DM)	ViT (CLIP)-based	♡, ♢	✗	Optional	DN7
	✗ (T2I DM)	Non.	♡	✗	✓	DN8
Visual to image	✗ (T2I DM)	Convolution-based	♣	✗	✗	DN9
	✗ (T2I DM)	ViT-based	♡	✗	✗	DN10
Image composition	✗ (T2I DM)	Convolution-based	♡	✓	✗	DN11
	✗ (T2I DM)	ViT (CLIP)-based	♡, ♢	✓	✗	DN12
Layout control	✗ (T2I DM)	ViT (CLIP)-based	♢	✗	✗	DN13

Stack of conditioning mechanisms for sampling process			
Task	**Backbone model**	**Conditioning mechanism**	**Serial Number**
Text-to-image	Uncond DM	Guidance	SP1
Image restoration	Conditional restoration DM	Revising Diffusion Process	SP2
	Uncond DM	Revising Diffusion Process	SP3
	Uncond DM	Guidance	SP4
	Uncond DM	Conditional Correction	SP5
Image editing	Uncond DM / T2I DM	Inversion	SP6
	T2I DM	Inversion, Conditional Correction	SP7
	T2I DM	Inversion, Attention Manipulation	SP8
	T2I DM	Inversion, Attention Manipulation, Guidance	SP9
Visual to image	T2I DM	Guidance	SP10
Image composition	Uncond DM	Noise Blending	SP11
Layout control	T2I DM	Attention Manipulation	SP12
	T2I DM	Attention Manipulation, Guidance	SP13
General purpose	Unspecified	Noise Composition	SP14
	Unspecified	Classifier-free Guidance	SP15
	Unspecified	Universal Guidance Framework	SP16

image formulated as conditional distribution $p_{\mathcal{T}}(\mathbf{x}|\mathbf{c})$, we classify representative conditional image synthesis tasks into seven categories as shown in Fig. 1: (a) *Text-to-image* synthesizes images in accordance with text prompts, (b) *Image restoration* recovers clean images from their degraded counterparts, (c) *visual signal to image* converts given visual signals such as sketch, depth and human pose into corresponding images,

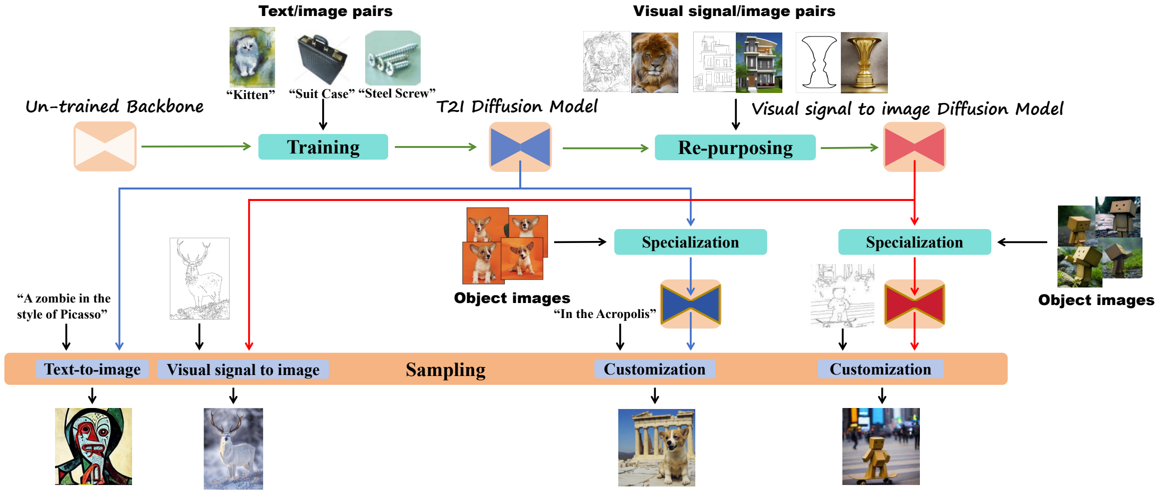

Figure 2: An example of the workflow to build denoising network via training, re-purposing and specialization stages for target conditional synthesis tasks. In this framework, a text-to-image (T2I) denoising network is firstly obtained via supervised learning on text/image pairs in *training stage*. Subsequently, this T2I denoising network is fine-tuned on visual signal/image pairs for visual signal to image task in *re-purposing stage*. Next, both T2I and visual signal to image denoising networks can be further fine-tuned on given object image in *specialization stage* to perform customization on the user-specified personal object. Figures are cited from Rombach et al. (2022); Zhang et al. (2023b); Ruiz et al. (2023); Li et al. (2023a).

(d) *Image editing* edits the given source images with provided semantic, structure or style information, (e) *Customization* creates different editing renditions for personal object specified by given images, (f) *Image composition* composes the objects and background specified in different images into a single image, and (g) *Layout control* controls the layout grounding of synthesized images with provided spatial information of foreground objects and background. We have sorted out the associations between various conditional synthesis tasks and conditioning mechanisms of representative existing works in Tab. 1.

3 Condition integration in denoising networks

The denoising network is the crucial component in the diffusion model (DM)-based synthesis framework, which estimates the noise added in each forward step to reverse the initial Gaussian noise distribution back into the data distribution. In practice, the most straightforward way to achieve conditional control in DM-based synthesis framework is incorporating the conditional inputs into the denoising network. In this section, we divide the condition integration in denoising network into three stages: (a) *training stage*: training a denoising network on paired conditional input and target image from scratch, (b) *re-purposing stage*: re-purposing a pre-trained denoising network to conditional synthesis scenarios beyond the task it was trained on, (c) *specialization stage*: performing testing-time adjustments on denoising network based on user-specified conditional input. Fig. 2 provides an examplar workflow to build desired denoising network for conditional synthesis tasks including text-to-image, visual signals to image and customization via these three condition integration stages. Next, we first review the fundamental conditional DMs modeled in *training stage* in Sec. 3.1. We then summarize the architecture design choices and condition injection approaches in *re-purposing stage* in Sec. 3.2. Finally, we introduce the works performing condition integration in *specialization stage* in Sec. 3.3. Fig. 3 illustrates the taxonomy proposed in this section.

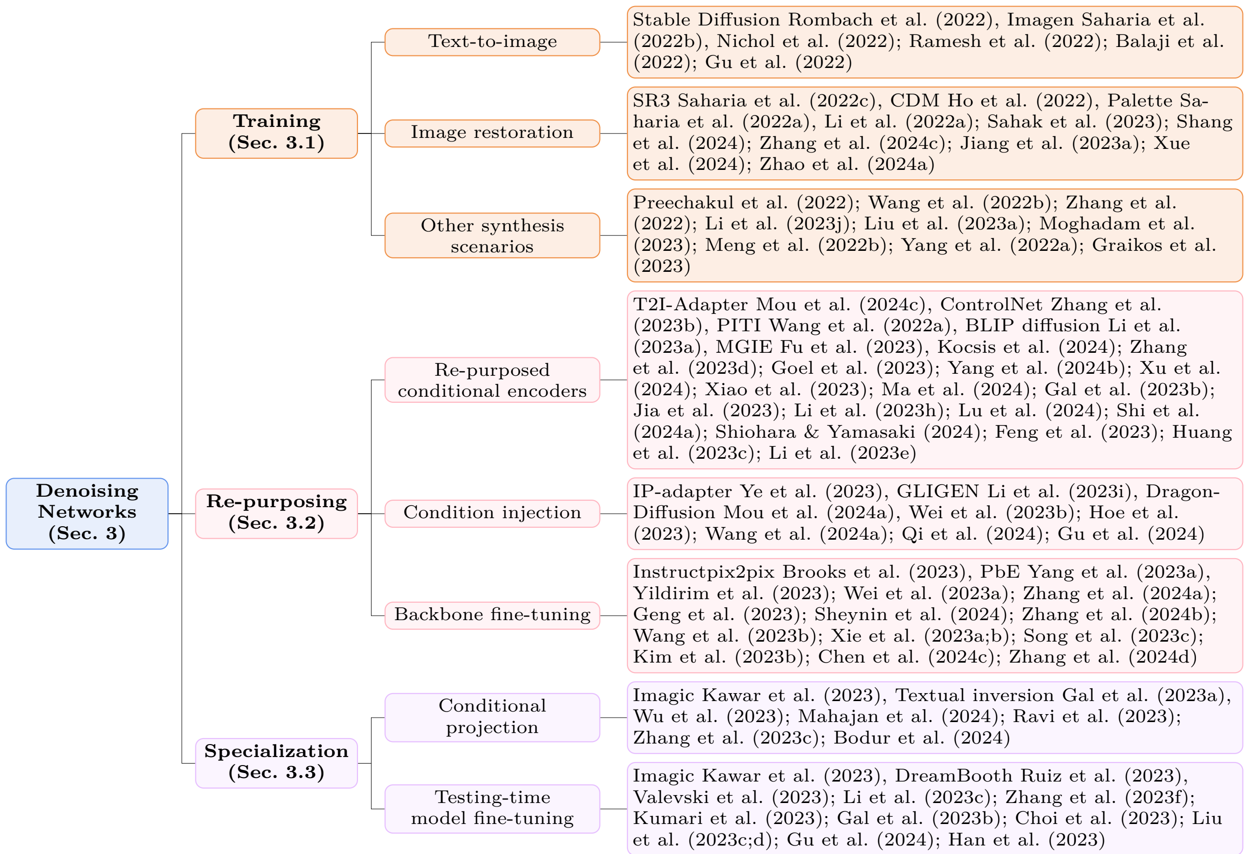

Figure 3: The proposed taxonomy of DCIS works performing condition integration in denoising network.

3.1 Condition Integration in the Training Stage

The most straightforward way to integrate the conditional control signal $\mathbf{c}$ into the denoising network is performing supervised training from scratch with the following loss function:

$$\mathbb{E}_{\mathbf{c},\mathbf{x}\sim p(\mathbf{x}|\mathbf{c}),\boldsymbol{\epsilon},t}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t,t,\mathbf{c}\right)\right\|_2^2\right], \tag{6}$$

where $\mathbf{c}$ and $\mathbf{x}$ denote the paired conditional inputs and target image. Thereby, the learned conditional denoising network $\boldsymbol{\epsilon_\theta}\left(\mathbf{x}_t,t,\mathbf{c}\right)$ can be employed to sample from $p(\mathbf{x}|\mathbf{c})$.

Next, we introduce the existing conditional denoising networks trained from scratch, focusing their model architectures, conditioning mechanisms, which are crucial for creating the connection between the conditional inputs and its corresponding image. Because of the conditioning architectures and mechanisms are designed based on the target scenarios, we categorize these works based on the their applications, represented by text-to-image and image restoration.

3.1.1 Conditional Models for Text-to-Image (T2I)

Text-to-image is a fundamental task in the field of conditional image synthesis, which establishes the connection between images and the semantic space of text descriptions. Because of the expressiveness of the text semantic space, text-to-image DMs always serve as the *backbone* for more complicated conditional synthesis tasks including image editing Kawar et al. (2023); Hertz et al. (2023); Brooks et al. (2023), customization Gal et al. (2023a); Ruiz et al. (2023), visual signal to image Mou et al. (2024c); Zhang et al. (2023b), image composition Yang et al. (2023a) and layout control Wang et al. (2024a); Li et al. (2023i).

The main challenge in modeling a effective text-to-image framework lies in (a) precisely capture the users' intention described in text prompts and (b) build the connection between text and image in acceptable computational cost. In practice, DM-based text-to-image works design different text encoders base on Transformer encoder Nichol et al. (2022); Rombach et al. (2022), CLIP Ramesh et al. (2022); Balaji et al. (2022); Gu et al. (2022) or more powerful large language models Saharia et al. (2022b); Balaji et al. (2022) to extract the features from user provided text prompts. For computational efficiency, these works often train the DMs on a low-dimension space including compressed latent space Rombach et al. (2022); Gu et al. (2022) and low-resolution pixel space Nichol et al. (2022); Saharia et al. (2022c); Balaji et al. (2022); Ramesh et al. (2022), and subsequently enlarge the resolution of the synthesized results.

Next, we introduce representative text-to-image model: Stable Diffusion Rombach et al. (2022) and Imagen Saharia et al. (2022b), which serve as the *T2I backbone* for various conditional synthesis tasks.

Similar to VQ-VAE Van Den Oord et al. (2017) and VQ-GAN Esser et al. (2021), Stable Diffusion Rombach et al. (2022) employs a pre-trained autoencoder to compress the generative space into a low-dimensional latent space for computational efficiency. In the training stage, the text-conditioned diffusion model $\boldsymbol{\epsilon_\theta}(\mathbf{z}_t, t, \mathbf{c})$ is trained on this latent space to approximate the conditional distribution of the latent representations. In sampling process, the latent representation aligned with given text prompt is firstly generated by the conditional diffusion model on latent space, and then fed into the decoder to recover its corresponding high-quality image.

For conditional control, Stable Diffusion introduces a transformer text encoder to interpret the text prompt and convert into the text embedding. Subsequently text embedding is fused with the features in U-Net architecture of denoising network Rombach et al. (2022) via cross-attention mechanism. In practice, the encoder can be different domain-specific experts other than the text encoder. Thereby, Stable Diffusion can be employed into various conditional synthesis scenario beyond text-to-image.

Following up the pioneer DM-based text-to-image framework GLIDE Nichol et al. (2022) and Imagen Saharia et al. (2022b) prefer to train the conditional denoising network on a low-resolution image space and subsequently upsample the synthesized low-resolution image. In order to effectively capture the complexity and compositionality of arbitrary text prompts, Imagen employs pre-trained large language models (e.g., BERT Kenton & Toutanova (2019), GPT Radford et al. (2021), T5 Raffel et al. (2020)) as powerful text-encoders. For condition injection, Imagen Saharia et al. (2022b) concatenates the encoded text embedding to the key-value pairs of the self-attention layers in denoising network. In Imagen, the basic 64×64 text-to-image diffusion model is followed by two cascaded super-resolution diffusion models designed to enlarge the resolution of synthesized image from 64×64 to 1024×1024.

3.1.2 Conditional Models for Image Restoration

DM-based conditional training is also widely employed to recover the high-quality clean image $\mathbf{x}$ from a given degraded image $\mathbf{c}$ Saharia et al. (2022c;a); Ho et al. (2022); Shang et al. (2024); Zhao et al. (2024a). These works primarily revolve around identifying the task-related features in degraded image as conditional input for supervised training and recovering the clean image based on the model trained on these core features.

2.1) Conditioning on degraded images. The most straightforward modeling approach is directly conditioning the diffusion model on the given degraded image via channel concatenation. Pioneer DM-based super-resolution method SR3 Saharia et al. (2022c) concatenates the low-quality reference image with the latent variable in the channel space of U-Net architecture. This simple operation empowers the U-Net architecture to comprehensively capture information in low-resolution image. Concurrent SRdiff Li et al. (2022a) shifts the generative space of SR3 to the residual space, and models the residuals between paired high and low resolution image to avoid regenerating the structures already existing in the low-resolution image. As a result, SRdiff performs on par with SR3 with significantly fewer computations. To adapt SR3 to real world restoration tasks, SR3+ Sahak et al. (2023) employs second-order degradation simulation to create real-world clean and degraded image-pairs to enhance the training dataset. Based on SR3 Saharia et al. (2022c), CDM Ho et al. (2022) proposes to cascade super-resolution DMs to enlarge image resolution, and Palette Saharia et al. (2022a) extends to more diverse image restoration tasks via supervised training on corresponding paired clean/degraded image datasets.

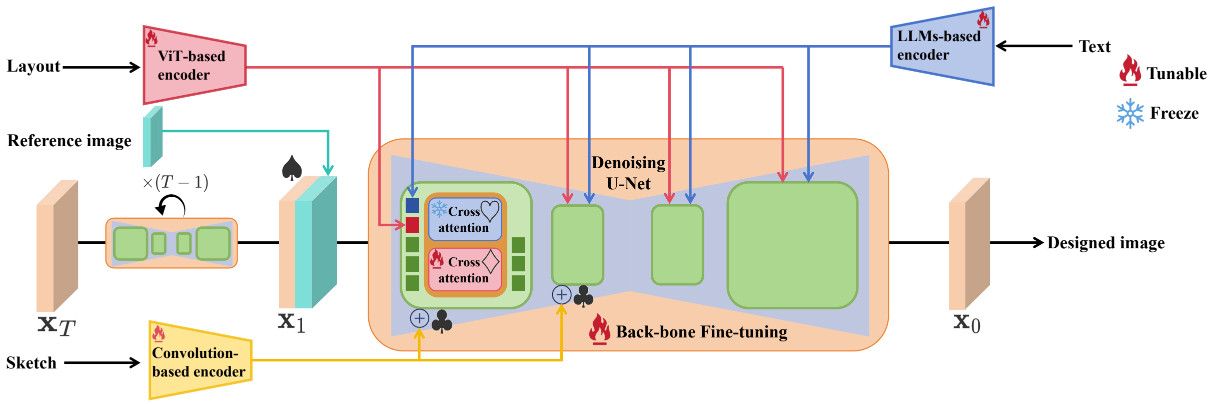

Figure 4: An illustration of the re-purposed denoising network based on text-to-image backbone, where ♠, ♡, ♣, ♢ denotes condition integration via channel-wise concatenation, T2I attention layers, addition and developed attention modules respectively as describes in Sec. 3.2.2.

2.2) Conditioning on pre-processed features. However, simply concatenating the degraded image in the channel space places a burden on the denoising network to extract information relevant to the restoration task from the unprocessed degraded image. To dedicate most modeling capacity on the task task-related features, a branch of restoration works Shang et al. (2024); Zhao et al. (2024a); Jiang et al. (2023a); Xue et al. (2024); Zhang et al. (2024c) prefer to firstly extract these features from the degraded image and subsequently conditioning the model on these task task-related features.

State-of-the-art super-resolution framework Resdiff Shang et al. (2024) employs a pre-trained CNN to generate a higher quality intermediate image for the initial degraded image, and conditions the denoising network on the intermediate image and its high-frequency details to synthesize the residual between intermediate image and clean image. For more complex restoration tasks including underwater image restoration Zhao et al. (2024a) and low-light image enhancement Jiang et al. (2023a); Xue et al. (2024), in which the given degraded image is severely corrupted, a branch of works prefer to condition the model on frequency information extracted by discrete wavelet transformations. To restore real-world text images under severe degradation, DiffTSR Zhang et al. (2024c) conducts parallel diffusion processes consist of an image diffusion model for image restoration and a text diffusion model for text recognition and employs a multi-modality module to interact the information of text and image diffusion process.

3.1.3 Conditional Models for Other Synthesis Scenarios

Although the mainstream DM-based frameworks for complicated conditional synthesis scenarios are established by re-purposing the text-to-image backbone, some works also prefer supervised training from scratch for different conditional synthesis tasks. Part of these works are early studies before the popularity of DM-based text-to-image models designed for tasks including image editing Preechakul et al. (2022) and visual signal to image Wang et al. (2022b); Zhang et al. (2022). Another part of these works are designed for novel or highly specialized tasks conditional synthesis scenarios including medical image synthesis Li et al. (2023j); Liu et al. (2023a); Moghadam et al. (2023); Meng et al. (2022b), graph-to-image Yang et al. (2022a) and satellite image synthesis Graikos et al. (2023), in which the conditional control signals are difficult to be aligned with the semantic space of the text-to-image backbone.

3.2 Condition Integration in the Re-purposing Stage

Currently, diffusion models (DMs) are employed in increasingly diverse and complex conditional synthesis scenarios Ye et al. (2023); Zhang et al. (2023b); Li et al. (2023e); Zhang et al. (2023d); Li et al. (2023i); Wang et al. (2024a); Shi et al. (2024b). Simply training denoising networks from scratch for each conditional synthesis scenario would place a heavy burden on computational resources. Fortunately, pre-trained text-to-image (T2I) DMs effectively associate text embedding with its corresponding image, which serves as a semantic powerful backbone for a wide range of conditional synthesis tasks beyond the T2I. Studies design task-specific denoising network based on T2I backbone and performing fine-tuning on paired conditional inputs and image to re-purpose the T2I-based denoising network to target task. In practice, the re-purposed denoising network can be divided into three key modules: (a) *Conditional encoder*: The module to encode the task-specific conditional inputs into feature embedding, (b) *Conditioning injection*: The module to inject task-related feature embedding into T2I backbone, (c) *Backbone*: The T2I backbone that can stay frozen or be fine-tuned during the re-purposing stage. In the re-purposing stage, conditional fine-tuning can be performed in each of these components for condition integration. Subsequently, we will summarize the design choice for these modules among current works performing condition integration in the re-purposing stage.

3.2.1 Re-purposed Conditional Encoders

In a T2I model, the text embedding is extracted from given text prompt through a text encoder and subsequently injected into the U-Net architecture through cross attention. To re-purpose the T2I backbone to tasks beyond text-to-image, various task-specific conditional encoders are designed to extract the features from conditional control signals other than text.

1.1) Convolutional layer-based encoder for visual signals. For visual signals, conditional encoders are mainly designed base on convolutional downsample blocks to extract multi-scale structure features.

Pioneer work T2I-Adapter Mou et al. (2024c) employs a four-layer convolutional network as a lightweight adapter to encode the visual signal into a set of multiscale features. ControlNet Zhang et al. (2023b) provides a more powerful architecture as tje encoder for visual signals, which cloned the deep encoding layers from the U-Net architecture in Stable Diffusion. This ControlNet encoder inherits a wealth of prior knowledge in the Stable Diffusion backbone and serves as a deep, robust, and strong architecture for diverse visual signals. Currently, ControlNet delivers state-of-the-art results in diverse visual signal to image tasks and becomes a the widely-employed conditional encoder various more complicated conditional synthesis scenarios including explicit lighting control Kocsis et al. (2024), image composition Zhang et al. (2023d), image editing Goel et al. (2023); Zhang et al. (2024d) and virtual try-on Kim et al. (2024); Zeng et al. (2024).

1.2) ViT-based encoder for images. In practice, Vision Transformer (ViT)-based encoders are widely employed to extract the features of conditional control signals in the form of images. Generally, visual signals can also be viewed in the form of image, the pioneering work PITI Wang et al. (2022a) designs a ViT-based encoder to map given visual signal into its corresponding text embedding for the T2I backbone. ImageBrush Yang et al. (2024b) also employs a ViT-based encoder to extract the visual editing instruction described by paired images before/after editing. Prompt-free Diffusion Xu et al. (2024) employs a more powerful Context Encoder (SeeCoder) based on SWIM-L Liu et al. (2021) to convert image into meaningful visual embedding. For customization, a branch of works Xiao et al. (2023); Ma et al. (2024); Shi et al. (2024a); Gal et al. (2023b); Jia et al. (2023); Li et al. (2023h); Lu et al. (2024); Li et al. (2023a); Shiohara & Yamasaki (2024) maps the given personal object into features on the textual space via different ViT-based image encoders designed on the framework of CLIP Radford et al. (2021), SWIN Liu et al. (2021), BLIP Li et al. (2023b) or ViT-based ArcFace encoder Deng et al. (2019).

1.3) LLMs-based encoder for image editing. In order to enhance the semantic information in the given text prompt, a branch of works prefer to design more powerful Large Language Models (LLMs)-base encoders for text-based image editing, Fu et al. (2023); Huang et al. (2023c); Li et al. (2023e) leverages a trainable Multimodal Large Language Models (MLLMs) Liu et al. (2024b) module as the encoder for the given source image and editing instruction. Ranni Feng et al. (2023) used LLMs to convert description or editing prompts into a semantic panel, which serves as an intermediate representation that contains rich structure and semantic information.

3.2.2 Condition Injection

In order to more effectively incorporate information from conditional inputs into the denoising network during the re-purposing stage across various conditional synthesis scenarios, studies in this field have designed different task-specific conditional injection approaches to handle different types of conditional control signals. Here, we categorize these methods into the following four categories.

2.1) Condition injection via concatenation ♠. For conditional inputs in form of image, a direct condition injection approach is following the concatenation strategy proposed by SR3 Saharia et al. (2022c), which concatenates the image form conditional inputs to the latent variable in the channel space of the U-Net architecture. In practice, this conditioning strategy is usually performed with backbone fine-tuning to handle conditional synthesis tasks that involve complex conditional inputs composed of multimodal components, including instruction-based editing Brooks et al. (2023); Sheynin et al. (2024); Geng et al. (2023) and image composition Zhang et al. (2023d); Song et al. (2023c); Xie et al. (2023a).

2.2) Condition injection via T2I attention layers ♡. In the T2I backbone, the cross-attention layers serve as the conditioning module to inject text embedding into the U-Net architecture. Currently, a branch of works also employ the cross-attention layers in T2I backbone to inject the features extracted from task-specific conditional encoders Wang et al. (2022a); Yang et al. (2024b); Xu et al. (2024); Xiao et al. (2023); Gal et al. (2023b); Jia et al. (2023); Li et al. (2023a); Shiohara & Yamasaki (2024); Zeng et al. (2024).

2.3) Condition injection via addition ♣. Because of the alignment between the architecture of conditional encoder and the U-Net encoder in T2I backbone, for convolutional layer-based encoders Mou et al. (2024c); Zhang et al. (2023b), the extracted features are injected via directly adding these features to the corresponding intermediates layers of U-Net architecture in T2I backbone.

2.4) Condition injection via developed attention modules ♢. To achieve more fine-grained control over the synthesized image, some works design developed task-specific attention modules for condition injection in target conditional synthesis scenarios Ye et al. (2023); Li et al. (2023i); Wei et al. (2023b); Wang et al. (2024a); Mou et al. (2024a). A branch of works prefer to incorporate extra attention module into the T2I backbone to inject the task-specific conditional control signals Ye et al. (2023); Wei et al. (2023b); Li et al. (2023i); Hoe et al. (2023); Wang et al. (2024a). IP-adapter Ye et al. (2023) employs additional image cross-attention layers to inject the image embedding into the T2I backbone. For customization, ELITE Wei et al. (2023b) leverages two parallel cross-attention layers to inject extracted global and local information of given personal object separately.

In T2I backbone, attention layers control the structure and layout information of synthesized image. To exert accurate object-level layout control, a branch of works prefer to add a trainable attention-module between self-attention and cross-attention layers Li et al. (2023i); Ma et al. (2024); Shi et al. (2024a); Hoe et al. (2023); Wang et al. (2024a). GLIGEN Li et al. (2023i) adds a gated self-attention layer to U-Net architecture to inject provided layout information. This conditioning strategy is further employed in customization works Ma et al. (2024); Shi et al. (2024a) to integrate patch features extracted from personal object images. To perform more detailed layout control, InteractDiffusion Hoe et al. (2023) designs an attention-based Human-Object Interaction module to inject the interactions between objects. InstanceDiffusion Wang et al. (2024a) projects different forms of object-level control signals including single points, scribbles, bounding boxes or intricate instance segmentation masks into the feature space through a UniFusion block, and inject these features with a Instance-Masked Attention module.

Another line of works modify the cross-attention mechanism in T2I backbone to achieve more precise control Qi et al. (2024); Mou et al. (2024a); Lu et al. (2024); Gu et al. (2024). Different from IP-adapter Ye et al. (2023), DEADiff Qi et al. (2024) concatenates the key and value features from image and text embedding respectively and perform a single fused cross-attention mechanism to achieve multimodal conditional control. In practice, performing fused attention mechanism to inject multimodal control signals along with text embedding is also employed in instruct-based editing Li et al. (2023f) and pose-guided person image synthesis Lu et al. (2024). To perform local control based on multiple regional prompts, Mix-and-show Gu et al. (2024) proposes an attention localization strategy in the re-purposing stage, which substitutes the attention map in specified regions with the attention map generated based on the regional prompts.

3.2.3 Backbone Fine-tuning

Currently, most of the re-purposing works confine the fine-tuning only on conditional encoders and condition injection modules to ease the computational burden. However, for conditional inputs contain multimodal components or intricate semantics, performing fine-tuning while freezing the parameters in T2I backbone often fails to fully understand intrinsic connections between the conditional input and target image. In these scenarios, fine-tuning the T2I backbone together with encoders and condition injection modules is a more preferable choice. Base on the fine-tuning strategy, we categorize these works into two types: (a) Fully supervised fine-tuning on annotated dataset, and (b) Self-supervised fine-tuning on bare image dataset.

3.1) Fully supervised fine-tuning on the annotated dataset. In practice, we can re-purpose the T2I backbone on the annotated dataset of paired conditional input and image in accordance with the specific task via fully supervised fine-tuning. For some synthesis tasks involving complex conditional inputs, a major difficulty lies in collecting sufficient training data to fine-tune the model Brooks et al. (2023); Zhang et al. (2023d). For instruct-based editing task which refers to using instruction instead of text description to guide the editing process, Instructpix2pix Brooks et al. (2023) provides an effective approach for automatically synthesizing training datasets. Firstly, InstructPix2Pix employs a fine-tuned GPT-3 Brown et al. (2020) to synthesize editing triplets composed of input captions, edit instructions and output captions. Subsequently, Instructpix2pix leverages Prompt-to-Prompt Hertz et al. (2023) to synthesize paired images corresponding to the input captions and output captions, which serves as the paired images before/after editing. This contribution leads to a line of works on DM-based instruction editing. A branch of follow-up works attempt to enhance the T2I backbone in some specific tasks by augmenting the training dataset for target scenario including object removal and inpainting Yildirim et al. (2023), global editing Li et al. (2023f), dialog-based editing Wei et al. (2023a), continuous editing Zhang et al. (2024a). InstructDiffusion Geng et al. (2023) and Emu-edit Sheynin et al. (2024) fine-tune the T2I backbone on larger and more comprehensive synthesized datasets for a wide range of vision tasks including image editing, segmentation, keypoint estimation, detection, and low-level vision. To achieve more accurate editing, Fu et al. (2023); Huang et al. (2023c); Li et al. (2023e) fine-tune the T2I backbone with a more powerful MLLMs-based conditional encoder to enhance the editing prompts. Based on reinforcement learning, HIVE Zhang et al. (2024b) fine-tunes the instruct-based editing model with a reward model reflecting the human feedback for editing performance.

3.2) Self-supervised fine-tuning on bare image dataset. In non-general conditional synthesis scenarios involving image composition or mask-based editing, the form of conditional inputs may be complicated. For example, a classic image composition task aims to fuse a foreground reference image into the background main image within the mask region. In these tasks, collecting annotated training data pairs is almost impossible. A feasible approach is to create paired data based on the target scenario through cropping on a bare image dataset, and thereby fine-tune the T2I backbone in a self-supervised manner. For image composition task, PbE Yang et al. (2023a) randomly crops the foreground objects from the source image as reference image and corresponding mask, while the remained background as the background main image. Subsequently, PbE Yang et al. (2023a) fine-tunes the T2I backbone with paired cropped reference image and main image. In practice, such strategy is widely employed in conditional synthesis scenarios involve inpainting Wang et al. (2023b); Xie et al. (2023a) and composition Song et al. (2023c); Kim et al. (2023b); Zhang et al. (2023d); Xie et al. (2023b); Chen et al. (2024c). To generate reasonable masks for text-based inpainting, Imagen Editor Wang et al. (2023b) employs an off-the-shelf object detector to generate mask on the image in captioned image datasets, which covers a region relevant to the text caption of image. SmartBrush Xie et al. (2023a) randomly augments the cropped training masks to create accurate instance masks, which facilitates the T2I backbone to follow the shape of the input mask at testing-time.

For image composition, the greatest challenge faced by the self-supervised fine-tuning strategy is how to avoid the trivial copy-and-paste solution caused by the training data cropped from a single image Yang et al. (2023a); Xie et al. (2023b); Zhang et al. (2024d). Currently, image composition frameworks resort compress the information in the conditional inputs into an information bottleneck. This, in turn, forces the T2I backbone to interpret the intrinsic connections between the conditional input and the desired image, thereby effectively avoiding the copy-and-paste solution. PbE Yang et al. (2023a) and Dreaminpainter Xie et al. (2023b) select part of the image tokens for condition injection to create information bottleneck. ObjectStitch Song et al. (2023c) employs a two-stage fine-tuning strategy to decouple the fine-tuning stages

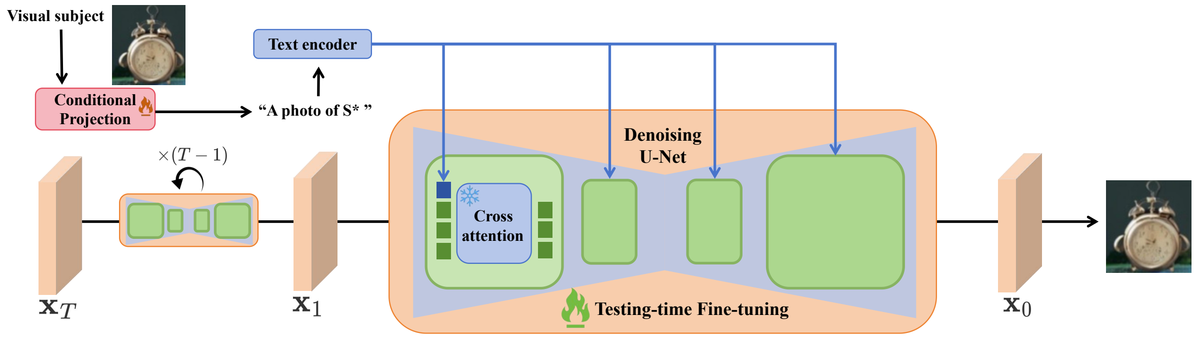

Figure 5: The specialization process to align a given personal object (the clock) with a pesudo-word S^* in the conditional space of a text-to-image backbone. The clock image is from Textual Inversion Gal et al. (2023a).

of the conditional encoder and the T2I backbone. Zhang et al. (2024d); Chen et al. (2024c); Zhang et al. (2023d) prefer to remove or mask out the information such as colors, textures or background in source image to prevent identical mapping.

3.3 Condition Integration in the Specialization Stage

Although theoretically we can incorporate any form of conditional inputs $\mathbf{c}$ into the denoising network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c})$ during the training and re-purposing stages, for complicated conditional synthesis scenarios, incorporating such control signals into the conditional space of denoising network faces challenges in collecting annotated training dataset and modeling the complicated correlation between conditional inputs and desire results. This limits the model capability to deal with zero-shot or few-shot conditional inputs.

A straightforward idea to remedy these issues is to align the given conditional inputs with the conditional space of a general T2I backbone through a specialization stage. As shown in Fig. 5, the specialization for given specific conditional inputs is typically achieved by (a) *conditional projection*, which projects the given conditional inputs onto the conditional space of the T2I backbone via embedding optimization Kawar et al. (2023); Gal et al. (2023a), or Vision-Language Pre-training (VLP) framework Li et al. (2022b; 2023b), (b) *testing-time model fine-tuning*, which fine-tunes the denoising network to insert the conditional inputs into the prior of the T2I backbone. In practice, works perform condition integration in specialization stage are mainly targeted to image editing and customization tasks to achieve desired edits on user-specified visual subjects including source image(image editing) and personal objects(customization) while preserving the characteristics and details in these visual subjects Kawar et al. (2023); Ruiz et al. (2023); Gal et al. (2023a).

3.3.1 Conditional Projection

To perform editing or customization tasks, a widely employed approach is projecting the given visual subject into corresponding text representation on the conditional space of text-to-image model.

1.1) Conditional embedding optimization. In order to find a proper text embedding for given visual subject, a branch of works directly search for the optimal embedding for the user-specified conditional inputs by optimizing the following objective function:

$$\mathbf{v}^* = \arg\min_{\mathbf{v}} \mathbb{E}_{\mathbf{x}=\mathbf{c}_I, \boldsymbol{\epsilon}, t} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{v})\|_2^2 \right], \tag{7}$$

where $\mathbf{v}^*$ denotes the optimized text embedding for the user-specified visual subject $\mathbf{c}_I$, and $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ denotes the T2I backbone. The embedding $\mathbf{v}_*$ serves as a pseudo-word S^* for the visual subject and can be further composed into various natural language prompts to create different editing renditions for given visual subject Kawar et al. (2023); Gal et al. (2023a).

For image editing, Imagic Kawar et al. (2023) optimizes the embedding $\mathbf{v}^*$ for the source image. Subsequently, Imagic perform interpolation between optimized source embedding $\mathbf{v}^*$ and target embedding $\mathbf{v}_{tgt}$ to obtain $\overline{\mathbf{v}} = \eta \cdot \mathbf{v}_{tgt} + (1-\eta) \cdot \mathbf{v}^*$, which serves as the conditional input for denoising process. Diffusion Disentanglement Wu et al. (2023) optimizes the time-specified combination weights $\lambda_{1:T}$ of the source and target text embedding along the sampling process instead of interpolation to retrieve time-adaptable embedding for editing. To reduce the computational cost of the optimization process, Zhang et al. (2023c); Mou et al. (2024b) first employed image encoder to generate a coarse embedding of the given visual subject, and subsequently fine-tuning the coarse embedding via optimization.

Pioneer customization work Textual inversion Gal et al. (2023a) perform optimization to discover the text embedding $\mathbf{v}^*$ for personal object described by a few reference images (typically 3 to 5). This optimized embedding $\mathbf{v}^*$ serves as the pseudo-pronoun S^* for the personal object in further conditional sampling process. To provide human-readable text description instead of text embedding for given personal object, PH2P Mahajan et al. (2024) employ quasi-newton L-BFGS Shanno (1970) to directly optimize discrete tokens from a existing pre-specified vocabulary for target image.

1.2) Employing VLP models. However, performing time-consuming optimization process for each new visual subject hinders the deployment of these methods in application scenarios. Therefore, a branch of works prefer to employ Vision-Language Pre-training (VLP) models to directly generate the embedding for given visual subjects Zhang et al. (2023c); Li et al. (2023a).

BLIP Li et al. (2022b) is a strong VLP framework to synthesize captions for given images, which is widely employed in image editing tasks to generate an initial text prompt to describe the uncaptioned source image Zhang et al. (2023c); Li et al. (2023a); Bodur et al. (2024); Parmar et al. (2023). BLIP can also be used to enhance user-provided prompts for eliminates editing failure caused by missing contexts in the coarse input prompts Kim et al. (2023c). Besides, PRedItOR Ravi et al. (2023) prefer to leverage DALL-E2 Ramesh et al. (2022) to fuse the source image with the target prompt by performing SDEdit Meng et al. (2022a) process on the CLIP embedding space.

3.3.2 Testing-time Model Fine-Tuning

In editing and customization tasks, simply employing the denoising network modeled in scenario-orient training and re-purposing stage always fails to retain the characteristics and details in the user-specified visual subject, due to the lack of prior knowledge Kumari et al. (2023). To customize the T2I backbone for user-specified conditional inputs, approaches in this category resort to perform testing-time fine-tuning on the T2I backbone to insert the given visual subjects into the denoising network Ruiz et al. (2023); Kumari et al. (2023).

To better preserve the outlook of source image in editing tasks, a branch of works Kawar et al. (2023); Valevski et al. (2023); Zhang et al. (2023c;f) represented by Imagic Kawar et al. (2023) fine-tune the T2I backbone to bind the source image with its corresponding text description $\boldsymbol{c}_{src}$ in the conditional space. In order to simultaneously editing the foreground and background in source image, LayerDiffusion employ Segment Anything Model (SAM) Kirillov et al. (2023) to create masks for foreground objects. Subsequently, LayerDiffusion Li et al. (2023c) fine-tunes the T2I backbone with a designed loss composed of the diffusion loss in both foreground and background region to editing the foreground object and background independently. SINE Zhang et al. (2023f) introduces a patch-based fine-tuning strategy which incorporates the positional embedding into conditional T2I space to synthesize arbitrary-resolution edited image.

For the customization task, DreamBooth Ruiz et al. (2023) fine-tunes the T2I backbone to entangle a fixed unique identifier with the semantic meaning of the personal object. To alleviate the computational burden in the testing-time fine-tuning, followed up works Kumari et al. (2023); Gal et al. (2023b); Choi et al. (2023); Liu et al. (2023c;d); Gu et al. (2024); Han et al. (2023) prefer to only fine-tune a specific part of model parameters. CustomDiffusion Kumari et al. (2023) fine-tunes only the cross-attention layers. E4T Gal et al. (2023b) optimizes low-rank adaptations (LoRA) Hu et al. (2021) of weight residuals in cross- and self-attention layers to further reduce computational cost. Cones Liu et al. (2023c) fine-tunes the attention layer concept neurons highly-related to the given visual subject. Cones2 Liu et al. (2023d) and Mix-and-show

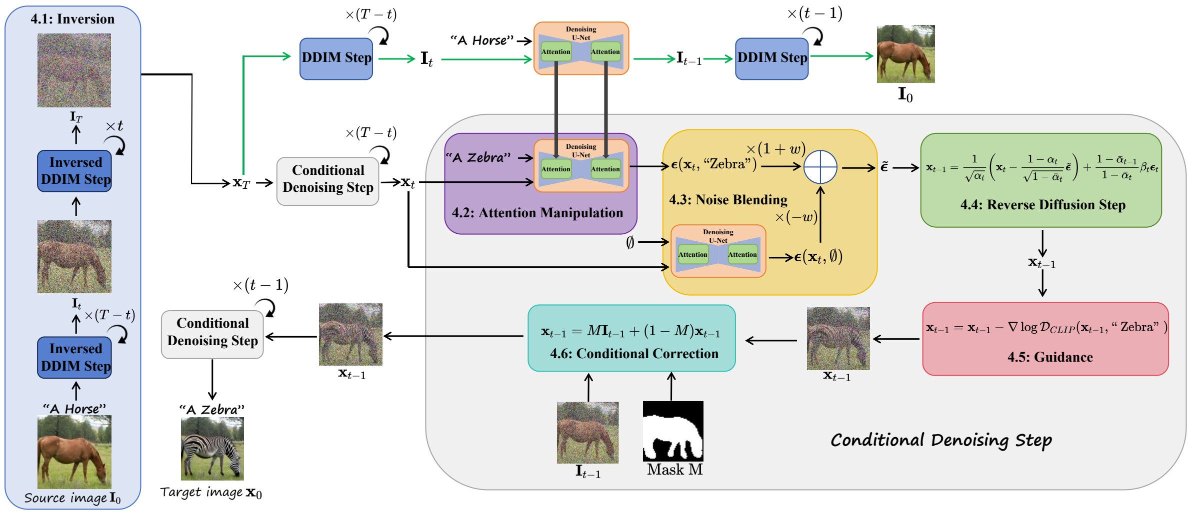

Figure 6: An example of the conditional sampling process for image editing, in which we incorporate all six mainstream in-sampling conditioning mechanisms for diffusion sampling process to provide a comprehensive overview of the content in this section. The sample images are from Diffedit Couairon et al. (2023).

Gu et al. (2024) resort to fine-tune the text encoder in T2I backbone. SVDiff Han et al. (2023) fine-tunes the singular values of the decomposed convolution kernels.

4 Condition integration in the sampling process

In DM-based image synthesis frameworks, the sampling process iteratively reserve noisy latent variable into desired image with the prediction of the denoising network. As mentioned in Sec. 3, integrating the conditional control signals into the denoising network always requires time-consuming training, fine-tuning or optimization. To ease the burden for conditioning the denoising network, numerous works perform condition integration in the sampling process to ensure the consistency between synthesized image and given conditional input without computational intensive supervised-training or fine-tuning Su et al. (2023); Hertz et al. (2023); Liu et al. (2022); Kawar et al. (2022); Dhariwal & Nichol (2021); Choi et al. (2021).

Based on how the conditional control signals are incorporated into the sampling process, we divide mainstream in-sampling conditioning mechanisms into six categories: (a) *inversion*, (b) *attention manipulation*, (c) *noise blending*, (d) *revising diffusion process*, (e) *guidance* and (f) *conditional correction*. We illustrate these conditioning mechanisms with an exemplary image editing process in Fig. 6. In this section, we will introduce the core idea of these conditioning mechanisms and summarize the corresponding representative works as taxonomized in Fig. 7.

4.1 Inversion

In diffusion model (DM)-based image synthesis, the starting latent variable controls the spatial structure and semantics of synthesized result. Inversion process provides an effective way to encode the given source image back into its corresponding starting latent variable and effectively preserve the image structure and semantics for further editing. In this section, we firstly summarize the inversion approaches in Sec. 4.1.1. Next, we will discuss the applications of inversion in various conditional synthesis scenarios in Sec. 4.1.2.

4.1.1 Inversion Approaches

Mainstream inversion approaches perform inversion based on the forward diffusion process, deterministic sampling process, and stochastic sampling process. We denote these three basic inversion pathways as *noise-adding inversion*, *deterministic inversion*, and *stochastic inversion*, respectively. Due to accumulated errors

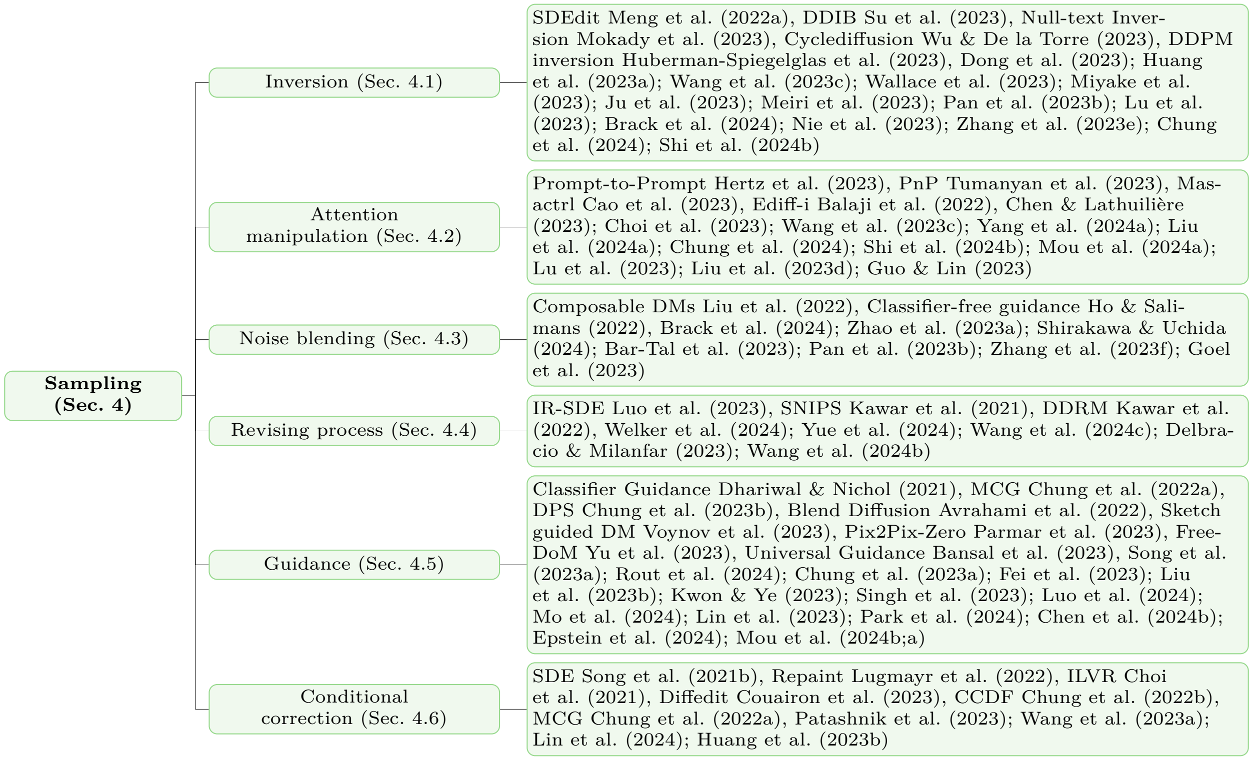

Figure 7: The proposed taxonomy of DCIS works performing condition integration in sampling process.

in the discrete diffusion process, the naive inversion process often fails to preserve details in the source image, especially with classifier-free guidance. Therefore, numerous works propose enhancements to these basic inversion approaches to ensure perfect reconstruction of the source image.

1.1) Noise-ddding inversion. Noise-Adding Inversion performs a standard forward diffusion process to inverse the source image to a certain noise step T', i.e., $q\left(\mathbf{x}_{T'} \mid \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_{T'} ; \sqrt{\bar{\alpha}_{T'}} \mathbf{x}_0,\left(1-\bar{\alpha}_{T'}\right) I\right)$, where the latent variable $\mathbf{x}_{T'}$ is a mixture of source image and Gaussian noise.

1.2) Deterministic inversion. However, noise-adding inversion may smooth out details in the source image. To more precisely preserve image features, deterministic inversion is proposed to encode the source image $\mathbf{x}_0$ into its corresponding latent variable $\mathbf{x}_T$ with the discretization of diffusion ODEs such as DDIM Song et al. (2021a). Theoretically, with a sufficiently large diffusion step T, DDIM inversion can guarantee perfect reconstruction, which ensures the latent variable $\mathbf{x}_T$ obtained from DDIM inversion to be a meaningful diffusion starting point encapsulating all features pertaining to the source image $\mathbf{x}_0$.

1.3) Stochastic inversion. However, DDIM inversion performs accurate inversion only when the diffusion time steps is sufficiently large, which always leads to unsatisfied results especially under classifier-free guidance. Therefore, a branch of works prefer to inverse the stochastic sampling process in Eq. 2. Different from the deterministic sampling process, which is determined by the starting point latent variable $\mathbf{x}_T$, the stochastic sampling process involves the noise vector $\boldsymbol{\epsilon}_t$ added in each reverse transition kernel. Therefore, we have to memorize each noise vector $\boldsymbol{\epsilon}_t$ along the inversion process to ensure the reconstruction property.

1.4) Enhanced inversion approaches. In conditional synthesis, the classifier-free guidance significantly magnified the accumulated error in inversion process, which leads to poor reconstruction and edit performance. Therefore, a series of inversion methods are developed to ensure the inversion performance under classifier-free guidance.

For deterministic inversion, some approaches prefer to *fine-tune* relevant parameters in the classifier-free guided sampling process to reduce the reconstruction error, including optimizing the null-text embedding Mokady et al. (2023), text embedding for the source image Dong et al. (2023), key and value matrix in the self-attention layers Huang et al. (2023a), and the prompt embedding for cross-attention layers Wang et al. (2023c). To get rid of the computational burden for fine-tuning, a branch of works has developed *tuning-free* approaches for perfect reconstruction Wallace et al. (2023); Han et al. (2024); Miyake et al. (2023); Ju et al. (2023). EDICT Wallace et al. (2023) achieves precise DDIM inversion by utilizing an equivalent reversible process consisting of two coupled noise vectors. Negative-prompt Inversion Miyake et al. (2023) demonstrates the prompt of the source image can serve as a training-free substitute for null-text embedding. Proxedit Han et al. (2024) further enhance the reconstruction performance of Negative-prompt Inversion Miyake et al. (2023) by incorporating a regularization term in classifier-free guidance to prevent over-amplifying the editing direction in sampling process. Fixed-point Inversion Meiri et al. (2023) and AIDI Pan et al. (2023b) perform fixed-point iterations in each step of DDIM inversion to reduce the accumulation errors due to the discrete DDIM process. Besides, Fixed-point Inversion Meiri et al. (2023) provides a brief cycle of fixed-point iterations for the VAE-encoded latent representation of source image to eliminate the misfit between latent representation and given text prompt in latent diffusion model. TF-ICON Lu et al. (2023) and LEDITS++ Brack et al. (2024) perform inversion based on high-order diffusion differential equation solvers Lu et al. (2022a;b) which significantly accelerates the inversion process and improve the accuracy of inversion.

For stochastic inversion, theoretically, any sampling sequence starts with source image can be employed as the iterative latent variables in stochastic inversion process. However, arbitrary sampling sequence will deviate from the prior marginal distribution of latent variables and harm the editing ability in reconstruction process. To construct a reasonable sampling sequence, pioneer work Cyclediffusion Wu & De la Torre (2023) firstly samples a $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and subsequently denoise it based on the source image $\mathbf{x}_0$ to recover the sampling sequence. DDPM inversion Huberman-Spiegelglas et al. (2023) constructs an editing-friendly sequence by sampling each intermediate latent variable $\mathbf{x}_t$ independently based on the source image $\mathbf{x}_0$ and reconstructs the source image up to noise precision to avoid error accumulation . SDE-Drag Nie et al. (2023) provides a theoretical fundamental to explain the superiority in editing performance of stochastic inversion comparing to deterministic inversion. It demonstrates that the KL-Divergence between the distribution of edited image and prior data distribution decrease in stochastic inversion while remaining in widely used deterministic inversion.

4.1.2 Applications of Inversion in Conditional Synthesis

Inversion process converts the provided source image into its corresponding latent variable. In practice, this latent variable can serve as the starting point for sampling process to perform basic image-to-image translation, text-based image editing or be further manipulated for more complicated tasks.

Image-to-image translation target to translate the content in a given source image into the desired appearance, which serves as the foundation for image editing. Pioneer SDEdit Meng et al. (2022a) translate a given out-of-domain source image into its counterpart in target domain by denoising the noise-adding inversed source image with the denoising network trained on target domain. This process preserves the content in source image while endowing it with appearance in the target domain.

Based on deterministic inversion, DDIB Su et al. (2023) introduces a highly flexible technique for image-to-image translation between two manifolds α and β via a simple process $\mathbf{x}^* = \mathcal{D}_\beta(\mathcal{E}_\alpha(\mathbf{x}))$, where $\mathbf{x}$ and $\mathbf{x}^*$ denote the source and target image on manifold α and β respectively, $\mathcal{E}_\alpha$ and $\mathcal{D}_\beta$ denote the deterministic inversion and sampling process performed with the diffusion models for manifold α and β. DDIB process can be performed with two independently trained diffusion models or a diffusion model conditioned on different control signals.

In practice, text-based editing task, which targets to edit the source image $\mathbf{c}_I$ described by $\mathbf{c}_{src}$ to align with target text prompt $\mathbf{c}_{tgt}$, can be achieved by performing the DDIB image-to-image translation process as $\mathbf{x}^* = \mathcal{D}_{\mathbf{c}_{tgt}}(\mathcal{E}_{\mathbf{c}_{src}}(\mathbf{c}_I))$, where $\mathbf{c}_I$, $\mathbf{x}^*$ are paired source and edited image, and $\mathcal{D}_{\mathbf{c}_{tgt}}$ and $\mathcal{E}_{\mathbf{c}_{src}}$ denotes the sampling process conditioned on target prompts and the inversion process conditioned on source prompts. However, this editing process can only roughly ensure the consistency in semantics and overall structure while always failing to precisely preserve the intricate details in source image. In order to more accurately recover the details in source image in editing process, inversion is always performed with other conditioning mechanisms in the editing process. Performing conditional correction with mask is a preferable choice to preserve the region not requiring editing Couairon et al. (2023); Li et al. (2023c); Patashnik et al. (2023); Yang et al. (2024a); Wang et al. (2023a); Lin et al. (2024); Huang et al. (2023b). Another choice is performing attention manipulation during the editing process to incorporate the outlook of source image, as discussed in Sec. 4.2 Hertz et al. (2023); Tumanyan et al. (2023). Besides, a branch of works employ model fine-tuning in specialization stage or conditional projection described in Sec. 3.3 to inject the detailed outlook of source image into the T2I backbone Kawar et al. (2023); Zhang et al. (2023c).

Besides, based on the task-specific conditional encoders to convert multi-model conditional inputs into text embedding, this inversion-based editing process can also be employed in conditional synthesis tasks beyond text-based editing. For example, InST Zhang et al. (2023e) denoise the noisy reference image obtained by noise-adding inversion with the denoising network conditioned on the embedding vectors extracted from the style image to achieve style transfer editing.

For more complicated conditional synthesis scenarios, the latent variable obtained from inversion can be manipulated to incorporate additional information beyond the source image. For image composition, a branch of works prefer to fuse the latent variable obtained from inversion process for different source images Chung et al. (2024); Lu et al. (2023). Style Injection in Diffusion Chung et al. (2024) fuse the latent variable of both style and content image obtained by DDIM inversion to perform style transfer. TF-ICON Lu et al. (2023) compose the inverted main and reference images for image compositing. In drag-based editing, we can adjust the corresponding area in the latent variable based on the provided drag instructions. Dragdiffusion Shi et al. (2024b) optimize the latent variable with designed motion supervision loss for drag-style manipulation. The stochastic inversion-based work, SDE-Drag Nie et al. (2023), manipulates the latent variable through a copy-and-paste strategy instead of performing optimization in the latent space.

4.2 Attention Manipulation

After determining the starting point for the sampling process via sampling from Gaussian distribution or inversion methods, the sampling process is performed by iterative denoising steps. As pointed out in E4T Gal et al. (2023b), the attention layers in the denoising network have the greatest influence on the predicted noise in each denoising step and thereby control the structure and layout of synthesized image. Therefore, a branch of works resort to design task-specific manipulation to the attention layers in denoising network to achieve more accurate control over the spatial layout and geometry Hertz et al. (2023); Tumanyan et al. (2023); Lu et al. (2023); Patashnik et al. (2023). Different from the works Li et al. (2023i); Ye et al. (2023) performing fine-tuning on modified attention module in re-purposing stage, approaches in this category manipulates the attention layers via tuning-free replacement or localization during sampling process.

4.2.1 Replacement Manipulation

Pioneer attention manipulation works are designed preserve the structure of source image during the inversion-based image editing process. Prompt-to-Prompt Hertz et al. (2023) performs parallel sampling processes for the inverted source image separately conditioned on source and target prompts. During the parallel sampling process, Prompt-to-Prompt replaces the cross-attention maps in editing branch with its counterpart in reconstruction branch in order to preserve the structure of source image during the editing sampling process. This replacement strategy is further employed in followed up works for face aging editing Chen & Lathuilière (2023) and customization-based editing Choi et al. (2023). P2Plus Wang et al. (2023c) further replaces the editing branch self-attention map in the unconditional noise predictor network with its counterpart in reconstruction branch to obtain more accurate editing capabilities with classifier-free guidance. In order to prevent undesired changes caused by cross-attention leakage, DPL Yang et al. (2024a) optimizes the word embedding corresponding to the noun words in source prompt to produce more suitable cross-attention maps for attention replacement.

PnP Tumanyan et al. (2023) points out that more detailed spatial features are restored in self-attention layers comparing to the cross-attention maps. Therefore, a branch of editing works Tumanyan et al. (2023); Liu et al. (2024a); Cao et al. (2023) prefer to replace query and key feature in self-attention layer to achieve better structure preservation. This replacement strategy is followed by works designed for drag-based editing

Shi et al. (2024b); Mou et al. (2024a) and style transfer Chung et al. (2024) to ensure the consistency between synthesized result and provided source image.

4.2.2 Attention Localization

To achieve more precise layout control for the synthesized image, a branch of works manipulate the attention layers with masks or segmentation indicating the locations of objects Patashnik et al. (2023); Lu et al. (2023); Balaji et al. (2022).

Some of these works propose localized self-attention mechanisms to address different regions separately and locate the contents into desired regions. Masactrl Cao et al. (2023) and Object-Shape Variation Patashnik et al. (2023) firstly extract the regions with attention value above a threshold in the cross-attention maps for object text tokens as foreground masks. Subsequently, Masactrl performs self-attention for foreground and background separately to prevent confusion between the foreground objects and the background. Object-Shape Variation Patashnik et al. (2023) restrict the region for attention replacement on the background not requiring editing instead of injecting the full self-attention maps in every denoising step. For image composition, TF-ICON Lu et al. (2023) fuses the attention features extracted from the reconstruction process for the reconstruction branches of both main and reference images via cross-attention mechanism to create a composite self-attention map seamlessly blending the two images.

Another line of works incorporate an increment into the cross-attention map to adjust the attention values in the region for designated objects and thereby achieve layout control for synthesized image. Pioneer text-to-image work Ediff-i Balaji et al. (2022) successfully guides the object described by the nouns in the text prompt to the specified area by enhancing the attention values in the corresponding region. Similarly, Cones2 Liu et al. (2023d) increases the attention values in the region corresponding to desired objects while reducing the attention values in irrelevant regions to perform layout control. For image editing, FoI Guo & Lin (2023) amplifies the attention value in the region of foreground object to be edited to achieve more precisely control for the objects in accordance with editing instructions.

4.3 Noise Blending

Noise blending process fuses noises predicted by different (conditional) DMs to perform single sampling process controlled by multiple conditional signals.

4.3.1 Noise Composition

In conditional synthesis scenarios aiming at synthesizing images conditioned on multiple control signals, directly training a denoising network to take all conditional inputs always leads to an unsustainable training cost. A widely employed approach to tackle these tasks is predicting the noise $\boldsymbol{\epsilon}_i$ for each conditional component $\mathbf{c}_i$ separately and subsequently composing these noise to acquire a novel proxy noise $\tilde{\boldsymbol{\epsilon}}$ controlled by all the conditional control signals without supervised-learning. Composable Diffusion Models Liu et al. (2022) present a noise composition approach based on Bayes' formula as follows to perform multi-conditional synthesis:

$$\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right) + \sum_{i=1}^{n} w_i \left(\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t, \mathbf{c}_i\right) - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)\right), \tag{8}$$

where the unconditional denoising network $\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)$ can be trained along with the conditional model by substituting the conditional parameter with emptyset $\emptyset$.

The noise composition can be performed based on masks or layouts to locate the objects in provided conditional inputs into desired regions. To perform image editing on multiple instructions, LEDITS++ Brack et al. (2024) calculates the mask for the region related to each instruction with the grounding information in cross-attention layers and noise estimations. Subsequently, LEDITS++ Brack et al. (2024) performs noise composition based on the formula of Eq.8 while restricting effect of the conditional term $\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t, \mathbf{c}_i\right) - \boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, t\right)$ of each editing instruction $\mathbf{c}_i$ in its corresponding mask region. In order to fuse the generated results of two diffusion models, MagicFusion Zhao et al. (2023a) firstly generates mask by contrasting the saliency map of the two diffusion models to differentiate the region controlled by each model. Subsequently, MagicFusion Zhao et al. (2023a) settles the noise into the region controlled by its corresponding diffusion model. Similarly, NoiseCollage Shirakawa & Uchida (2024) independently estimates the noises for each individual object and then merges them with a crop-and-merge operation based on the provided layouts. In order to perform more seamless noise composition, Multi-diffusion Bar-Tal et al. (2023) blends the noise by solving an optimization objective with closed-form optimal solution, which ensure the consistency of composed noise map $\tilde{\boldsymbol{\epsilon}}$.

4.3.2 Classifier-Free Guidance

As described in Sec.4.5, in traditional guidance, adjusting the guidance strength scaling factor w allows us to effectively balance the quality and diversity of synthesized samples. However, estimating the likelihood term $p_t(\mathbf{c} \mid \mathbf{x}_t)$ in traditional guidance is challenging.

Classifier-free guidance Ho & Salimans (2022) provides a new pathway to achieve balance the quality and diversity of synthesized samples without likelihood estimation, which can be achieved by performing extrapolation blending between the conditional noise prediction and the unconditional noise prediction as: $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c}) = (1+w)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$. In this formula, the parameter w controls the strength of guidance and the trade-off between sample quality and diversity. In practice, setting the scaling factor w to a value greater than zero can significantly enhance the sample quality and the consistency to the conditional control signal $\mathbf{c}$. In order to alleviate the negative impact of classifier-free guidance on sample diversity, followed works Sadat et al. (2024); Kynkäänniemi et al. (2024) propose dynamic classifier-free guidance, in which the guidance scaling factor w is reduced during the denoising process with high noise levels.

Moreover, some works also propose variations of classifier-free guidance for different conditional synthesis scenarios. Instructpix2pix Brooks et al. (2023) and Pair diffusion Goel et al. (2023) develop the classifier-free guidance to adjust the conditioning strength for each component in multiple conditional inputs by decomposing the multi-conditional score function. For customization tasks, SINE Zhang et al. (2023f) interpolates the noise prediction on specialized and pre-trained model to obtain conditional noise prediction in classifier-free guidance, which alleviates the overfitting in the specialized model. Null-text Guidance perturbs the classifier-free guidance by altering the noise-level in unconditional prediction to smooth out some realistic details and create cartoon-style images. For inversion-based editing, AIDI Pan et al. (2023b) proposes a blended classifier-free guidance based on the positive/negative masks indicating the area to be edited or preserved, which enables larger guidance scales and ensures more accurate editing results.

4.4 Revising Diffusion Process

Most of in-sampling conditioning mechanisms such as Guidance, Conditional Correction and Attention Manipulation performs modification on the standard formulation of the denoising step, which leads to deviations from the predetermined sampling trajectory and results in artifacts in synthesized images. Therefore, a branch of works prefer to incorporate the conditional control signals into the denoising step via revising the formulation of standard diffusion process to adapt the conditional synthesis task Luo et al. (2023); Yue et al. (2024); Kawar et al. (2022); Wang et al. (2024b). Thereby, the conditional control signals can be incorporated into the corresponding reverse diffusion step of the revised diffusion process without deviations from the diffusion formulation.

Based on the revision on diffusion process, these works can be divided into two categories: (a) *mean-reverting SDEs*, which revise the diffusion process to preserve the information in conditional inputs in image restoration, (b) *decomposition-based noise redefinition*, which incorporate a sequence of additive noises in the sampling process on spectral space to revise the noise-level mismatch in noisy linear problem.

4.4.1 Mean-Reverting SDEs

In numerous restoration tasks, most structure and semantic features of the target image is provided by the degraded image $\mathbf{c}$. To avoid consuming part of the model capability on regenerating these features from pure Gaussian noise, some studies design novel diffusion process in which the diffused output $\mathbf{x}_T$ approximates a noisy version of degraded image $\mathbf{c}$ instead of pure Gaussian noise Welker et al. (2024); Luo et al. (2023); Yue et al. (2024); Wang et al. (2024c); Delbracio & Milanfar (2023). IR-SDE Luo et al. (2023) construct a set of mean-reverting SDEs identified by degraded image $\mathbf{c}$, which models the diffusion process from clean image $\mathbf{x}$ to a Gaussian distribution averaged on degraded image. Subsequently, IR-SDE trains a conditional denoising network to predict the score function in the reversed mean-reverting SDEs to recover the clean image from the noisy degraded image. Similarly, ResShift Yue et al. (2024) and DriftRec Welker et al. (2024) construct an iterative degradation process from a high-resolution image to its corresponding low-resolution image as diffusion process and train a conditional denoising network to reverse the degradation process for super-resolution. SinSR Wang et al. (2024c) distills the sampling process of ResShift Yue et al. (2024), thereby achieving one-step DM-based super-resolution. InDI Delbracio & Milanfar (2023) constructs a continuous forward degradation process derived from interpolation: $\mathbf{x}_t = (1-t)\mathbf{x} + t\mathbf{c}$ and trains a denoising network on paired clean/degraded image to predict clean image $\mathbf{x}_0$ from latent variable $\mathbf{x}_t$. Subsequently, image restoration can be performed by reversing the interpolation-based degradation process with the prediction of this denoising network.

4.4.2 Decomposition-Based Noise Redefinition

This kind of methods construct novel diffusion process to recover image $\mathbf{x}$ from its partial measurement $\mathbf{c}$ in the noisy linear inverse problems as follows $\mathbf{c} = \boldsymbol{H}\mathbf{x} + \mathbf{n}$, where $\mathbf{H}$ is a known linear degradation matrix, $\mathbf{n} \sim \mathcal{N}\left(0, \sigma_{\mathbf{c}}^2 \mathbf{I}\right)$ is an i.i.d. additive Gaussian noise with known variance. In practice, numerous restoration tasks including inpainting, super-resolution, colorization can be written in form of this noisy linear inverse problems. SVD Decomposition-based methods firstly perform SVD decomposition on the linear degradation matrix $\mathbf{H}$ to decouples the components in the measurement $\mathbf{c}$. Thereby, the components in measurement $\mathbf{c}$ on spectral space can be viewed as a noisy version of their counterparts derived from clean image $\mathbf{x}$. In order to incorporate the measurement $\mathbf{c}$ into the diffusion process while preventing the mismatch in noise-level caused by the noise in measurement $\mathbf{c}$, decomposition-based methods design a proper noise sequence to link the noise in the measurement $\mathbf{c}$ with the noise added in the standard diffusion process. It can be proven that the optimized unconditional denoising network pre-trained on the prior of clean image $\mathbf{x}$ is also the optimal solution for the variational objective of the designed novel diffusion process. Thereby, we can perform sampling process in the spectral space to recover clean image $\mathbf{x}$ from its noisy counterpart $\mathbf{c}$ based on pre-trained unconditional denoising network. SNIPS Kawar et al. (2021) and DDRM Kawar et al. (2022) construct SVD decomposition-based novel diffusion process in spectral space based on the annealed Langevin dynamics framework provided by NCSN Song & Ermon (2019) and the Markov chain diffusion process provided by DDPM Ho et al. (2020) respectively.

Different from SNIPS and DDRM, DDNM Wang et al. (2024b) construct a general solution $\hat{\mathbf{x}}$ based on range-null space decomposition which holds $\mathbf{H}\hat{\mathbf{x}} \equiv \mathbf{c}$. In each denoising step, DDNM Wang et al. (2024b) project the denoising output $\mathbf{x}_{0|t}$ onto the general solution to guarantee the consistency between denoising output $\mathbf{x}_{0|t}$ and given measurement $\mathbf{c}$. For noisy linear inverse problem $\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{n}$, DDNM Wang et al. (2024b) incorporates a scaling factor into the formulation of general solution and designs noise sequence corresponding to the scaling factor during sampling process to assure the noise level in $\mathbf{x}_{t-1}$ aligned with the definiation of $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_0\right)$ for pre-trained unconditional denoising network.

4.5 Guidance

In the field of conditional image synthesis, an intuitive idea to sample from the conditional distribution $p(\mathbf{x}|\mathbf{c})$ is approximating the conditional score function $\nabla_{\mathbf{x}_t} \log p_t\left(\mathbf{x}_t \mid \mathbf{c}\right)$ with conditional denoising network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c})$. Guidance provides another pathway to approximate the conditional score function without time-consuming conditional training, since the conditional score function can be decomposed into an unconditional score function and the gradient of log likelihood as follows:

$$\nabla_{\mathbf{x}_t} \log p_t\left(\mathbf{x}_t \mid \mathbf{c}\right) = \nabla_{\mathbf{x}_t} \log p_t\left(\mathbf{c} \mid \mathbf{x}_t\right) + \nabla_{\mathbf{x}_t} \log p_t\left(\mathbf{x}_t\right) \tag{9}$$

where the score function $\nabla_{\mathbf{x}_t} \log p_t\left(\mathbf{x}_t\right)$ can be estimated by an unconditional denoising network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$. Guidance-based methods design task-specific guidance loss function to reflect the consistency between intermediate latent variable $\mathbf{x}_t$ and conditional inputs $\mathbf{c}$ at each time step t, which serves as the estimation for the log likelihood $\log p_t(\mathbf{c} \mid \mathbf{x}_t)$.

For multiple conditional inputs, guidance can also be employed to perform conditional control for part of the conditional inputs. In practice, we can split the conditional inputs $\mathbf{c}$ into components $\mathbf{c}_0$ and $\mathbf{c}_1$ which are incorporate into the diffusion synthesis framework with conditional denoising network and guidance respectively. In this case, the conditional score function can be written as $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{c}_0, \mathbf{c}_1) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{c}_1 \mid \mathbf{x}_t, \mathbf{c}_0) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{c}_0)$. In this formulation, $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{c}_0)$ can be estimated by a denoising network conditioned on $\mathbf{c}_0$ and the log likelihood $\log p_t(\mathbf{c}_1 \mid \mathbf{x}_t, \mathbf{c}_0)$ can be estimated with the guidance loss.

Currently, guidance-based methods are employed in a wide range of conditional synthesis scenarios with task-specific guidance loss. Subsequently, we categorize these approaches based on the target applications.

4.5.1 Classifier Guidance

The pioneer guidance work Classifier Guidance Dhariwal & Nichol (2021) trains an auxiliary classifier $p_\phi(\mathbf{c} \mid \mathbf{x}_t)$ as the guidance loss function for image synthesis conditioned on class label $\mathbf{c}$. However, for more complicated conditional control signal $\mathbf{c}$ beyond the class label, training an accurate classifier $p_\phi(\mathbf{c} \mid \mathbf{x}_t)$ is challenging. Therefore, followed up works designs more flexible guidance loss without training or optimizing to handle more complicate tasks.

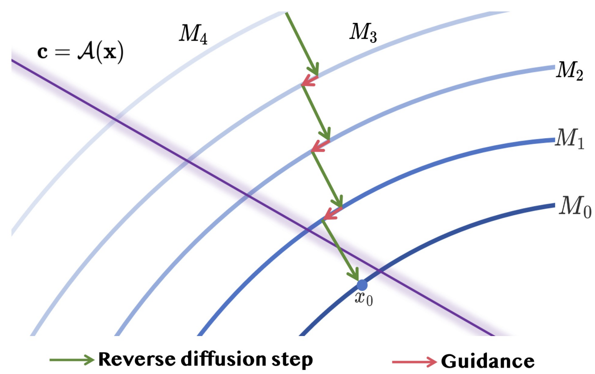

Figure 8: An illustration of the guided sampling process for inverse problems. The curve M_t denotes the data manifold of intermediate diffuse output $\mathbf{x}_t$. The guidance process (red arrow) moves $\mathbf{x}_t$ towards the data manifold satisfying the constrain $\mathbf{c} = \mathcal{A}(\mathbf{x})$, which is denoted as the purple line.

4.5.2 Guidance for Inverse Problems

As mentioned in Sec. 4.4.2, a wide range of restoration tasks can be expressed by recovering clean image $\mathbf{x}$ from a given partial measurement $\mathbf{c}$ in form of noisy inverse problem: $\mathbf{c} = \mathcal{A}(\mathbf{x}) + \boldsymbol{n}$, $\boldsymbol{n} \sim \mathcal{N}(0; \sigma_{\mathbf{c}}^2 \boldsymbol{I})$, where $\mathcal{A}$ is a known degradation function and $\boldsymbol{n}$ denotes the additive noise. In practice, approximating the likelihood $p_t(\mathbf{c}|\mathbf{x}_t)$ and perform guidance on sampling process is a widely employed strategy to solve noisy inverse problem. Fig. 8 provides an illustration of sampling process with guidance for inverse problem.

MCG Chung et al. (2022a) and DPS Chung et al. (2023b) approximate the gradient of likelihood as follows: $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{c} \mid \mathbf{x}_t) \approx \nabla_{\mathbf{x}_t} \log p(\mathbf{c}|\mathbf{x}_{0|t}) = -\frac{1}{\sigma_{\mathbf{c}}^2} \nabla_{\mathbf{x}_t} \left\| \mathbf{c} - \mathcal{A}\left(\mathbf{x}_{0|t}\right) \right\|_2^2$. The error of this estimation can be proven to converge to 0 as $\sigma_{\mathbf{c}} \to \infty$ in most inverse problems. IIGDM Song et al. (2023a) provides a more accurate estimation for the likelihood by approximating $p_t(\mathbf{x}_0 \mid \mathbf{x}_t)$ with a Gaussian distribution averaged on $\mathbf{x}_{0|t}$. In order to perform these guidance approaches for inverse problems on diffusion framework on latent space Rombach et al. (2022), PSLD Rout et al. (2024) adds an additional guidance term measuring the reconstruction ability of the intermediate denoising output $\mathbf{z}_{0|t}$ to avoid guiding the sampling trajectory towards latent variable $\mathbf{z}_0$ away from the manifold of real data.

However, these guidance approaches can only estimate the likelihood term in inverse problems with known concrete form of the degradation operator $\mathcal{A}(\cdot)$. This hinders the deployment of these approaches for unknown real world degradation. BlindDPS Chung et al. (2023a) explores the applicability of DPS to blind inverse problems, in which degradation operator $\mathcal{A}_{\boldsymbol{\varphi}}(\cdot)$ is parameterized with unknown parameter $\boldsymbol{\varphi}$. In order to identify the degradation parameter along with the sampling process for desired image, BlindDPS trains a diffusion model for the parameter $\boldsymbol{\varphi}$ in degradation operator. In sampling process, BlindDPS employed the similar approximation strategy as DPS Chung et al. (2023b) to estimate the likelihood term as follows:

$$p_t(\mathbf{c} \mid \mathbf{x}_t, \boldsymbol{\varphi}_t) \approx p\left(\mathbf{c} \mid \mathbf{x}_{0|t}, \boldsymbol{\varphi}_{0|t}\right). \tag{10}$$

Subsequently, BlindDPS performs parallel sampling process to simultaneously recover the clean image $\mathbf{x}$ and the unknown degradation parameter $\boldsymbol{\varphi}$ from conditional distribution $p(\mathbf{x}, \boldsymbol{\varphi}|\mathbf{c})$ with the estimated likelihood in Eq.10.

GDP Fei et al. (2023) offers a heuristic approximation for the likelihood term, which consists of a distance metric measuring the consistency to conditional inputs and a optional quality enhancement loss to control some desired properties in synthesized results. GDP can also be employed in blind inverse problems by optimizing the degradation parameters in degradation function $\mathcal{A}$ with the distance metric during sampling process.

4.5.3 Guidance for Semantic Control

Guidance can also be employed to ensure the consistency of diffused output and provided semantic control signals including text prompts or semantic images without time-consuming fine-tuning or training. In practice, semantic guidance loss is usually designed based on pre-trained CLIP model which learned a rich shared embedding space for image and text.

Blend Diffusion Avrahami et al. (2022) is the pioneer work in the field of semantic guidance, which targets to inpaint the masked region $\mathbf{c}_m$ in source image $\mathbf{c}_\mathbf{I}$ according to the provided text description $\mathbf{c}_d$. Blend Diffusion designs a CLIP guidance loss for the conditional inputs $\mathbf{c} = (\mathbf{c}_m, \mathbf{c}_I, \mathbf{c}_d)$ as follows:

$$L(\mathbf{x}_t, \mathbf{c}) = \mathcal{D}_{CLIP}\left(\mathbf{x}_{0|t}, \mathbf{c}\right) + \lambda \mathcal{D}_{bg}\left(\mathbf{x}_{0|t}, \mathbf{c}\right), \tag{11}$$

where $\mathcal{D}_{CLIP}$ measures the CLIP distance between the intermediate denoising output $\mathbf{x}_{0|t}$ and text description $\mathbf{c}_d$ in mask region for semantic-level alignment, and $\mathcal{D}_{bg}$ calculates the MSE and LPIPS similarity between $\mathbf{x}_{0|t}$ and source image $\mathbf{c}_\mathbf{I}$ in unmasked region for the faithfulness to source image.

In order to control the sampling process with both provided text prompt and style reference image, SDG Liu et al. (2023b) employs a linear combination of the CLIP distance from current denoising output to both text embedding and reference image embedding as the guidance loss. DiffuseIT Kwon & Ye (2023) introduce a more comprehensive guidance loss to perform image editing in accordance with given text prompt or style reference image. In addition to the CLIP distance, DiffuseIT also incorporates a structure loss calculated based on the self-attention features of the source image extracted from the Vision Transformer (ViT) to better preserve the structure of the source image.

4.5.4 Guidance for Visual Signals

In practice, a branch of works employ guidance to control the consistency between diffuse output and given visual signal. In order to measure the consistency between intermediate diffuse output and provided visual signal, some works train neural networks to project the intermediate diffuse output $\mathbf{x}_t$ onto its corresponding visual signal and leverage distance metric as the guidance loss for sketch-to-image Voynov et al. (2023) and stroke-to-image Singh et al. (2023). Readout Guidance Luo et al. (2024) provide a unified guidance-based framework for diverse visual signal to image task by training various readout heads to synthesize different task-specific visual feature maps reflecting the spatial layout or inherent correspondence in images to perform guidance. Different from these works, FreeControl Mo et al. (2024) prefers to impose guidance loss on the difference in the space of PCA components of self-attention map between the intermediate diffuse output and visual signal.

4.5.5 Guidance for Attention Layers

In DM-based conditional image synthesis, the attention layers in denoising network effectively control the layout, structure and semantics of synthesized image. However, directly manipulating the attention layers through replacement or localization as described in Section 4.2 introduces artificial modifications to the internal parameters of the denoising network and may impair its modeling capability. Therefore, a branch of works employ guidance to achieve softly control for attention layers.

For image editing, attention guidance is performed as substitution of attention replacement to softly control the consistency between source image and edited result. Pix2Pix-Zero Parmar et al. (2023) employs a guidance loss measuring the L_2 distance between the cross-attention maps in editing branch and reconstruction branch instead of the replacement manipulation in Prompt-to-prompt Hertz et al. (2023). In order to find a more expressive attention map as guide reference, Rediffuser Lin et al. (2023) employs a sliding fusion strategy to fuse the cross-attention maps obtained from sampling branches conditioned on source prompt, target prompt and an intermediate representation. EBMs Park et al. (2024) employs a energy function to guide the integration of the semantic information in editorial prompts with the structure and layout of source image restored in cross-attention layers.

Attention guidance can also be employed to perform attention localization. For object-level layout control, Chen et.al Chen et al. (2024b) employs guidance to control the cross-attention map, which locates the objects in text prompts into their desired bounding boxes. Self-guidance Epstein et al. (2024) extracts the various characteristics including position, size, shape and appearance of the desired object from the intermediate activations and attention maps. Subsequently, Self-guidance places constraints on these characteristics with guidance loss measuring their consistency to desired conditional control signal. For drag-based editing tasks which target to move certain foreground contents in source image into target region, Dragondiffusion Mou et al. (2024a) designs energy functions based on the cosine distance between intermediate features in the U-Net decoder as guidance to ensure correspondence between the original content region and target dragging region. DiffEditor Mou et al. (2024b) develops the guidance framework of DragonDiffusion Mou et al. (2024a) by introducing SDE-based sampling process on the masked region instead of ODEs to improve editing flexibility.

4.5.6 Enhanced Guidance framework

In some complicated conditional synthesis scenarios, simply incorporating the gradient of guidance loss in each denoising step may lead to artifacts and strange behaviors because of the failure in balancing the realness and guidance constraint satisfaction in guided sampling process. Therefore, some state-of-the-art guidance works provide enhanced unified guidance frameworks to more effectively fuse the prior knowledge in pre-trained model and the information in control signals. FreeDoM Yu et al. (2023) employs a time-travel strategy that rolls back the intermediate latent variable $\mathbf{x}_t$ to a certain previous time step $\mathbf{x}_{t+j}$ and resamples it to time step t again. This strategy inserts additional steps into the guided sampling process, allowing for a more seamless integration of the information from the pre-trained model and the conditional control signals. In order to enhance the consistency to conditional control signals, Universal Guidance Bansal et al. (2023) performs an m-step gradient descent optimization process to find the point with minimum guidance loss in the vicinity of the intermediate denoising output $\mathbf{x}_{0|t}$. Subsequently, this point is employed to infer the next latent variable $\mathbf{x}_{t-1}$.

4.6 Conditional Correction

In some conditional synthesis scenarios, the synthesized images are controlled by the constrains specified by conditional inputs $\mathbf{c}$ (such as the formulation of inverse problems). To ensure the synthesized result to be consistent to the inputs $\mathbf{c}$, conditional correction-based methods perform a correction operator on the intermediate diffuse output $\mathbf{x}_t$ (or $\mathbf{x}_{0|t}$), which directly projects the current diffuse output onto the data manifold satisfying the constrain imposed by given conditional control signal $\mathbf{c}$. Subsequently, this corrected latent variable will be pass into next denoising step. Fig. 9 provides an illustration of sampling process with conditional correction for inverse problem.

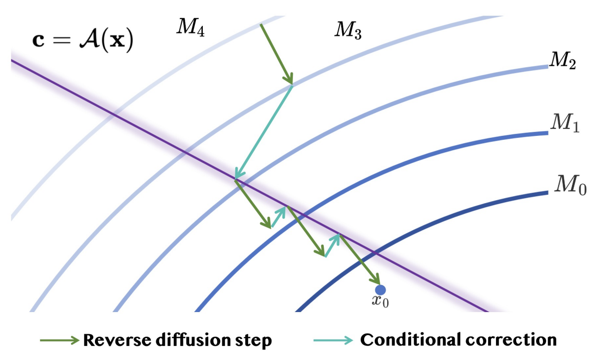

Figure 9: An illustration of the sampling process with conditional correction for inverse problem. The conditional correction process (cyan arrow) projects $\mathbf{x}_t$ onto the data manifold satisfying the constrain $\mathbf{c} = \mathcal{A}(\mathbf{x})$.

Currently, conditional correction are widely employed in image inpainting tasks, which involves synthesizing content for the masked region $\mathbf{c}_m$ in incomplete reference image $\mathbf{c}_y$. The constrain in inpainting tasks can be expressed as: $\mathbf{c}_y = (1 - \mathbf{c}_m) \odot \mathbf{x}$. Pioneer diffusion work SDE Song et al. (2021b) performs inpainting based on conditional correction by replacing the unmask region in denoising output $\mathbf{x}_{0|t}$ with its counterpart in reference image $\mathbf{c}_y$ to ensure the faithfulness to the content in unmasked region. Different from SDE Song et al. (2021b), Repaint Lugmayr et al. (2022) prefers to perform replacement correction on latent variable $\mathbf{x}_t$. Besides, Repaint rolls back the intermediate latent variable $\mathbf{x}_t$ to the previous time step and resamples it to time step t several times to eliminate the artifacts caused by conditional correction. The constrain in Super-resolution task can be written as: $\mathbf{c} = \phi_N \mathbf{x}$, where $\mathbf{c}$ denotes the low-resolution image of $\mathbf{x}$ downsampled by degradation matrix ϕ_N with factor N. ILVR Choi et al. (2021) performs conditional correction by substituting the low-frequency components in latent variable with its counterpart noisy low-resolution image to the consistency between degraded latent variable and its counterpart noisy reference low-resolution image.

Conditional correction are also widely employed in image editing tasks to preserve the background not requiring editing Couairon et al. (2023); Patashnik et al. (2023); Wang et al. (2023a); Lin et al. (2024); Huang et al. (2023b). With the provided mask for background in source image, text-based image editing tasks can be viewed as performing image inpainting for the foreground region based on given text prompt. However, the provided mask for background is always not available in editing tasks. Therefore, a branch of works propose approaches to generate masks or segmentation automatically by inferring the reasonable layout for the user-desired edited image based on the given source image and text prompt. Diffedit Couairon et al. (2023) identifies the mask for background by comparing differences in the denoising outputs of noisy source image conditioned on source prompt and target prompt. Object-Shape Variation Patashnik et al. (2023) segments the provided source image by the aggregating the attention map into clusters corresponding to different semantic segments and identifying the segments with the nouns in the text prompt based on the similarity between the segments and the cross-attention map of noun tokens. Besides, a branch of works Wang et al. (2023a); Lin et al. (2024); Huang et al. (2023b) employ pre-trained image segmentation modules to automatically generate masks or segmentation according to the structure information in the given source image and text prompt.

CCDF Chung et al. (2022b) proposes a general conditional correction formula for constrains in form of general noisy linear inverse problem. In practice, the conditional correction operator in Song et al. (2021b); Lugmayr et al. (2022); Choi et al. (2021) can be expressed in the general form provided by CCDF. Besides, CCDF provides a theoretical basis for the faithfulness of this corrected sampling trajectory to original sampling process. CCDF proves when the linear degradation operator $\boldsymbol{H}$ is a non-expansive mapping, the upper bound of the deviation in final output $\mathbf{x}_0$ will converge to a constant as the total diffusion step $T \to \infty$. MCG Chung et al. (2022a) further performs guidance on conditional correction framework provided by CCDF, which alleviates the deviation from original sampling process caused by conditional correction.

5 Challenges and Future Directions

Although DM-based conditional image synthesis has made remarkable progress in generating high-quality images aligned with various user-provided conditions, there remains a significant disparity between academic advancements and practical needs for conditional image synthesis. In this section, we summarize several main challenges in this field and identify potential solutions to address them in the future.

5.1 Sampling Acceleration

The time-consuming sampling process often creates a bottleneck of diffusion-based image synthesis, and its acceleration will facilitate the model deployment in practice Li et al. (2024b); Zhao et al. (2023b). Early works on sampling acceleration are devoted to reducing the number of sampling steps with better numerical solvers Song et al. (2021a); Lu et al. (2022a;b); Zhou et al. (2024); Chen et al. (2024a) or distilling pre-trained diffusion models to build short-cuts that enable faster sampling Salimans & Ho (2022); Meng et al. (2023); Song et al. (2023b); Chen et al. (2023). However, too few denoising steps with the distilled model may compromise the effectiveness of in-sampling condition integration. One feasible solution is to first train a model to approximate the conditional denoising outputs along the sampling process equipped with in-sampling conditioning mechanisms, and then perform distillation on this model Meng et al. (2023). Another important type of existing works reduces the computational cost of each denoising step by decreasing model parameters using techniques such as knowledge distillation Chen et al. (2021; 2022) and architecture search Li et al. (2024b); Kim et al. (2023a); Zhao et al. (2023b). Most of DM-based parameter compression approaches are currently tailored for text-to-image models. Analyzing whether the parameter redundancy also exists for models of other conditional synthesis tasks, similar to those in text-to-image models, and extending these model compression methods to more complicated downstream tasks, is another promising future direction.

5.2 Artifacts Caused by In-sampling Conditioning Mechanisms

In-sampling condition mechanisms summarized in Sec. 4 allows for flexible condition integration in DM-based image synthesis without performing time-consuming condition integration for the denoising network. However, these conditioning mechanisms introduce modification to the standard sampling process in diffusion framework and lead to deviations from the modeled data distribution, which resulting in artifacts in synthesized images Parmar et al. (2023); Lugmayr et al. (2022); Bansal et al. (2023); Yu et al. (2023). The vast majority of works resort to complex adjustment mechanisms to address the artifact issue caused by in-sampling condition integration. This includes time-step rolling back for guidance Yu et al. (2023), localization for attention map Cao et al. (2023); Lu et al. (2023) and diffusion process revision for restoration tasks Luo et al. (2023); Kawar et al. (2022). However, these methods are highly customized based on specific application scenarios. A feasible future direction for developing more generic solution is to perform lightweight fine-tuning on the denoising network with the diffusion loss based on the intermediate latent variables in the sampling process equipped with in-sampling conditioning mechanisms. This tends to smooth out artifacts under in-sampling conditioning mechanisms and synthesize desire images in a lower computational cost comparing to perform condition integration in denoising network .

5.3 Training Datasets

Among the various conditioning mechanisms, the most fundamental and effective pathway for condition integration is still the supervised learning on pairs of conditional input and image. Although training datasets are relatively sufficient for conditional synthesis tasks involving single modality conditional inputs, such as text-to-image Schuhmann et al. (2021; 2022), restoration Agustsson & Timofte (2017); Nah et al. (2017); Karras et al. (2019), and visual signal to image Lin et al. (2014); Caesar et al. (2018); Zhou et al. (2017), gathering enough data for tasks with complex, multi-modal conditional inputs like image editing, customization, and composition remains challenging. With the advancement of training and efficient fine-tuning techniques for large language models, various types of large models are constantly being developed with powerful multi-modal representation learning Brown et al. (2020); Li et al. (2022b; 2023b) and content generation abilities Hertz et al. (2023); Tumanyan et al. (2023), making it possible to leverage these pre-trained models to automatically produce desired training datasets. We may also consider self-supervised or weakly supervised learning to reduce the demand for a large amount of high-quality training data Zhang et al. (2023d); Xie et al. (2023b); Zhang et al. (2024d).

5.4 Robustness

Due to the lack of objective task-specific evaluation datasets and metrics in some complex tasks, studies for these tasks prefer to compare models based on a set of self-defined conditional inputs, making the performance appear overly optimistic. In fact, many renowned text-to-image models Ramesh et al. (2022); Saharia et al. (2022b); Rombach et al. (2022) have been found to produce unsatisfactory synthesized results for certain specific categories of text prompts, as demonstrated by the shortcomings of Imagen Saharia et al. (2022b) in generating facial images. Training dataset augmentation, carefully designed architecture of conditional encoders, and improved conditioning formulation for fine-grained control are promising directions for enhancing robustness.

Here we point out some pathways to address issues of robustness. First, for conditional inputs where the model performs poorly, augmenting the training dataset is a direct approach. Second, the difficulties to handle conditional inputs in a certain category may be due to the insufficient capability or unsuitability of the conditional encoder with this category of data. In this case, incorporating encoder architectures tailored for this data category into the conditional encoder, or designing more capable compound conditional encoders, becomes a preferable choice. Besides, performing specialization for given conditional inputs is also an effective pathway to provide robust results at the cost of time-consuming fine-tuning or optimization. Finally, employ sampling process conditioning mechanisms, such as guidance, conditional correction and attention manipulation, to achieve more detailed control can also prevent undesired synthesis results.

5.5 Safety

The developments in AI-generated content (AIGC) propelled by the superior performance of diffusion-based conditional synthesis and their downstream applications lead to severe safety concerns in aspects of bias and fairness, copyright, and the risk of exposure to harmful content. Safety-oriented DM-based conditional image synthesis is dedicated to mitigating these issues by embedding watermarks that are easily reproducible in DM-generated images to detect copyright infringement Yuan et al. (2024); Cui et al. (2023); Wen et al. (2023), and reducing bias by increasing model's orientation towards minority groups in basic unconditional or text-conditioned synthesis via classic conditioning mechanisms, such as fine-tuning Shen et al. (2023), guidance Um et al. (2024), and conditional correction Li et al. (2024a). Efforts have also been made in preventing harmful contents in the text-to-image task via harmful prompt detection Rombach et al. (2022), prompt engineering Li et al. (2024a) and safety guidance Schramowski et al. (2023). The current safety-focused efforts mainly concentrate on basic unconditional or text-conditioned synthesis. We believe that for more complex conditional synthesis scenarios, safety-oriented efforts in this area can be focused on four main aspects: (a) detecting harmful conditional inputs, (b) filtering and removing bias from the training dataset, (c) providing safety-focused guidance for the sampling process, and (d) implementing safety-focused fine-tuning of the denoising network.

6 Conclusion

This survey presents a thorough investigation of DM-based conditional image synthesis, focusing on framework-level construction and common design choices behind various conditional image synthesis problems across seven representative categories of tasks. Despite the progress made, efforts are still needed in the future to handle challenges in practical applications. Future researches should focus on gathering and creating sufficient high-quality and unbiased task-specific datasets, carefully designed conditional encoder architectures and in-sampling conditioning mechanisms for effective and robust conditional modeling to synthesize stable and flawless results. Trade-off between fast sampling and synthesization quality and is also a key issue for practical deployment. Finally, as a popular AIGC technology, it is necessary to fully consider the safety issues and legitimacy it brings.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020.

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR*, 2017.

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

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv:2211.01324*, 2022.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *CVPR*, 2023.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023.

Rumeysa Bodur, Erhan Gundogdu, Binod Bhattarai, Tae-Kyun Kim, Michael Donoser, and Loris Bazzani. iedit: Localised text-guided image editing with weak supervision. In *CVPR*, 2024.

Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *CVPR*, 2024.

Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. In *ICLR*, 2017.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023.

Shang Chai, Liansheng Zhuang, and Fengying Yan. Layoutdm: Transformer-based diffusion model for layout generation. In *CVPR*, 2023.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *AAAI*, 2021.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *CVPR*, 2022.

Defang Chen, Zhenyu Zhou, Jian-Ping Mei, Chunhua Shen, Chun Chen, and Can Wang. A geometric perspective on diffusion models. *arXiv:2305.19947*, 2023.

Defang Chen, Zhenyu Zhou, Can Wang, Chunhua Shen, and Siwei Lyu. On the trajectory regularity of ode-based diffusion sampling. In *ICML*, 2024a.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *WACV*, 2024b.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024c.

Xiangyi Chen and Stéphane Lathuilière. Face aging via diffusion-based editing. In *BMVC*, 2023.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021.

Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv:2305.15779*, 2023.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *NeurIPS*, 2022a.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, 2022b.

Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *CVPR*, 2023a.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023b.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *CVPR*, 2024.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2023.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE TPAMI*, 2023.

Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv:2306.04642*, 2023.

Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *TMLR*, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*, 2021.

Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *ICCV*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. In *NeurIPS*, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, 2023.

Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. *arXiv:2311.17002*, 2023.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv:2309.17102*, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023a.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG*, 2023b.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, et al. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv:2309.03895*, 2023.

Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv:2303.17546*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Alexandros Graikos, Srikar Yellapragada, Minh-Quan Le, Saarthak Kapse, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Learned representation-guided diffusion models for large-image generation. *arXiv:2312.07330*, 2023.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2024.

Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. *arXiv:2312.10113*, 2023.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *ICCV*, 2023.

Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *WACV*, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022.

Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. Interactdiffusion: Interaction control in text-to-image diffusion models. *arXiv:2312.05849*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021.

Jiancheng Huang, Yifan Liu, Jin Qin, and Shifeng Chen. Kv inversion: Kv embeddings learning for text-conditioned real image action editing. In *PRCV*, 2023a.

Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. Region-aware diffusion for zero-shot text-driven image editing. *arXiv:2302.11797*, 2023b.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv:2402.17525*, 2024.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. *arXiv:2312.06739*, 2023c.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. *arXiv:2304.06140*, 2023.

Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv:2304.02642*, 2023.

Hai Jiang, Ao Luo, Haoqiang Fan, Songchen Han, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM TOG*, 2023a.

Zeyinzi Jiang, Chaojie Mao, Yulin Pan, Zhen Han, and Jingfeng Zhang. Scedit: Efficient and controllable image diffusion generation via skip connection editing. *arXiv:2312.11392*, 2023b.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv:2310.01506*, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. In *NeurIPS*, 2021.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *NeurIPS*, 2022.

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 *CVPR*, 2023.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: A lightweight, fast, and cheap version of stable diffusion. *arXiv:2305.15798*, 2023a.

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, 2024.

Kangyeol Kim, Sunghyun Park, Junsoo Lee, and Jaegul Choo. Reference-based image composition with sketch via structure-aware diffusion model. *arXiv:2304.09748*, 2023b.

Sunwoo Kim, Wooseok Jang, Hyunsu Kim, Junho Kim, Yunjey Choi, Seungryong Kim, and Gayeong Lee. User-friendly image editing with minimal text input: Leveraging captioning and injection techniques. *arXiv:2306.02717*, 2023c.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nießner, and Yannick Hold-Geoffroy. Lightit: Illumination modeling and control for diffusion models. In *CVPR*, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. In *ICLR*, 2023.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv:2404.07724*, 2024.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022.

Hyunsoo Lee, Minsoo Kang, and Bohyung Han. Conditional score guidance for text-driven image-to-image translation. In *NeurIPS*, 2024.

Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023a.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *CVPR*, 2024a.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023b.

Pengzhi Li, Qinxuan Huang, Yikang Ding, and Zhiheng Li. Layerdiffusion: Layered controlled image editing with diffusion models. In *SIGGRAPH Asia*. 2023c.

Ruijun Li, Weihua Li, Yi Yang, Hanyu Wei, Jianhua Jiang, and Quan Bai. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation. *NCA*, 2023d.

Shufan Li, Harkanwar Singh, and Aditya Grover. Instructany2pix: Flexible visual editing via multimodal instruction following. *arXiv:2312.06738*, 2023e.

Sijia Li, Chen Chen, and Haonan Lu. Moecontroller: Instruction-based arbitrary image manipulation with mixture-of-expert controllers. *arXiv:2309.04372*, 2023f.

Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv:2308.09388*, 2023g.

Xiu Li, Michael Kampffmeyer, Xin Dong, Zhenyu Xie, Feida Zhu, Haoye Dong, Xiaodan Liang, et al. Warpdiffusion: Efficient diffusion model for high-fidelity virtual try-on. *arXiv:2312.03667*, 2023h.

Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. In *NeurIPS*, 2024b.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: open-set grounded text-to-image generation. In *CVPR*, 2023i.

Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *IEEE Trans. Med. Imaging*, 2023j.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-driven image editing via learnable regions. In *CVPR*, 2024.

Yupei Lin, Sen Zhang, Xiaojun Yang, Xiao Wang, and Yukai Shi. Regeneration learning of diffusion models with rich prompts for zero-shot image translation. *arXiv:2305.04651*, 2023.

Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *NeurIPS*, 2021.

Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *CVPR*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024b.

Jiaming Liu, Rushil Anirudh, Jayaraman J Thiagarajan, Stewart He, K Aditya Mohan, Ulugbek S Kamilov, and Hyojin Kim. Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In *ICCV*, 2023a.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: concept neurons in diffusion models for customized generation. In *ICML*, 2023c.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *NeurIPS*, 2023d.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv:2211.01095*, 2022b.

Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *ICCV*, 2023.

Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *CVPR*, 2024.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *CVPR*, 2024.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Image restoration with mean-reverting stochastic differential equations. In *ICML*, 2023.

Siwei Lyu. Interpretation and generalization of score matching. In *UAI*, 2009.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH*, 2024.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. *NeurIPS*, 2019.

Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *CVPR*, 2024.

Barak Meiri, Dvir Samuel, Nir Darshan, Gal Chechik, Shai Avidan, and Rami Ben-Ari. Fixed-point inversion for text-to-image diffusion models. *arXiv:2312.12540*, 2023.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022a.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023.

Xiangxi Meng, Yuning Gu, Yongsheng Pan, Nizhuan Wang, Peng Xue, Mengkang Lu, Xuming He, Yiqiang Zhan, and Dinggang Shen. A novel unified conditional score-based generative framework for multi-modal medical image completion. *arXiv:2207.03430*, 2022b.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv:2305.16807*, 2023.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *CVPR*, 2024.

Puria Azadi Moghadam, Sanne Van Dalen, Karina C Martin, Jochen Lennerz, Stephen Yip, Hossein Farahani, and Ali Bashashati. A morphology focused diffusion probabilistic model for synthesis of histopathology images. In *WACV*, 2023.

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

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *ICLR*, 2024a.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *CVPR*, 2024b.

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 *AAAI*, 2024c.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.

Shen Nie, Hanzhong Allan Guo, Cheng Lu, Yuhao Zhou, Chenyu Zheng, and Chongxuan Li. The blessing of randomness: Sde beats ode in general diffusion-based image editing. In *ICLR*, 2023.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Shaoyan Pan, Tonghe Wang, Richard LJ Qiu, Marian Axente, Chih-Wei Chang, Junbo Peng, Ashish B Patel, Joseph Shelton, Sagar A Patel, Justin Roper, et al. 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine & Biology*, 2023a.

Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. Effective real image editing with accelerated iterative diffusion inversion. In *ICCV*, 2023b.

Geon Yeong Park, Jeongsol Kim, Beomsu Kim, Sang Wan Lee, and Jong Chul Ye. Energy-based cross attention for bayesian context update in text-to-image diffusion models. In *NeurIPS*, 2024.

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.

Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *ICCV*, 2023.

Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit H Bermano, Eric Ryan Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. State of the art on diffusion models for visual computing. *arXiv:2310.07204*, 2023.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022.

Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *CVPR*, 2024.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: a unified diffusion model for controllable visual generation in the wild. In *NeurIPS*, 2023.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*, 2022.

Lingmin Ran, Xiaodong Cun, Jia-Wei Liu, Rui Zhao, Song Zijie, Xintao Wang, Jussi Keppo, and Mike Zheng Shou. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *CVPR*, 2024.

Hareesh Ravi, Sachin Kelkar, Midhun Harikumar, and Ajinkya Kale. Preditor: Text guided image editing with diffusion prior. *arXiv:2302.07979*, 2023.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *NeurIPS*, 2024.

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

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: unleashing the diversity of diffusion models through condition-annealed sampling. In *ICLR*, 2024.

Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *arXiv:2302.07864*, 2023.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022b.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2022c.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, 2023.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*, 2021.

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

Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai Zhang, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. In *AAAI*, 2024.

David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 1970.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. In *ICLR*, 2023.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. knn-diffusion: Image generation via large-scale retrieval. In *ICLR*, 2023.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *CVPR*, 2024.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024a.

Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *CVPR*, 2024b.

Kaede Shiohara and Toshihiko Yamasaki. Face2diffusion for fast and editable face personalization. In *CVPR*, 2024.

Takahiro Shirakawa and Seiichi Uchida. Noisecollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In *CVPR*, 2024.

Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv:2406.14555*, 2024.

Jaskirat Singh, Stephen Gould, and Liang Zheng. High-fidelity guided image synthesis with latent diffusion models. In *CVPR*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *ICLR*, 2023a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023b.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *CVPR*, 2023c.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *ICLR*, 2023.

Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv:2205.16007*, 2022.

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *CVPR*, 2024.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.

Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *ICLR*, 2024.

Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning a diffusion model on a single image. *ACM TOG*, 2023.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 2011.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH*, 2023.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *CVPR*, 2023.

Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *arXiv:2305.18047*, 2023a.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*, 2023b.

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv:2205.12952*, 2022a.

Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv:2207.00050*, 2022b.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pp. 1905–1914, 2021.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, 2024a.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2024b.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, 2024c.

Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *ICCV*, 2023c.

Jingxuan Wei, Shiyu Wu, Xin Jiang, and Yequan Wang. Dialogpaint: A dialog-based image editing model. *arXiv:2303.10073*, 2023a.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *CVPR*, 2023b.

Simon Welker, Henry N Chapman, and Timo Gerkmann. Driftrec: Adapting diffusion models to blind jpeg restoration. *IEEE TIP*, 2024.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv:2305.20030*, 2023.

Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, 2023.

Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *CVPR*, 2023.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv:2305.10431*, 2023.

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 2023a.

Shaoan Xie, Yang Zhao, Zhisheng Xiao, Kelvin CK Chan, Yandong Li, Yanwu Xu, Kun Zhang, and Tingbo Hou. Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models. *arXiv:2312.03771*, 2023b.

Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *CVPR*, 2024.

Minglong Xue, Jinhong He, Yanyi He, Zhipu Liu, Wenhai Wang, and Mingliang Zhou. Low-light image enhancement via clip-fourier guided wavelet diffusion. *arXiv:2401.03788*, 2024.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023a.

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

Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, and Ming-Hsuan Yang. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv:2211.11138*, 2022a.

Shiyuan Yang, Xiaodong Chen, and Jing Liao. Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *ACM MM*, 2023b.

Xiulong Yang, Sheng-Min Shih, Yinlin Fu, Xiaoting Zhao, and Shihao Ji. Your vit is secretly a hybrid discriminative-generative diffusion model. *arXiv:2208.07791*, 2022b.

Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike, et al. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. In *NeurIPS*, 2024b.

Zhen Yang, Ganggui Ding, Wen Wang, Hao Chen, Bohan Zhuang, and Chunhua Shen. Object-aware inversion and reassembly for image editing. In *ICLR*, 2023c.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv:2308.06721*, 2023.

Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv:2304.03246*, 2023.

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

Zihan Yuan, Li Li, Zichi Wang, and Xinpeng Zhang. Watermarking for stable diffusion models. *IEEE Internet of Things Journal*, 2024.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2024.

Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *CVPR*, 2024.

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv:2303.07909*, 2023a.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2024a.

Kaiduo Zhang, Muyi Sun, Jianxin Sun, Binghao Zhao, Kunbo Zhang, Zhenan Sun, and Tieniu Tan. Humandiffusion: a coarse-to-fine alignment diffusion framework for controllable text-driven person image generation. *arXiv:2211.06235*, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023b.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *ICLR*, 2023.

Shiwen Zhang, Shuai Xiao, and Weilin Huang. Forgedit: Text guided image editing via learning and forgetting. *arXiv:2309.10556*, 2023c.

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *CVPR*, 2024b.

Xin Zhang, Jiaxian Guo, Paul Yoo, Yutaka Matsuo, and Yusuke Iwasawa. Paste, inpaint and harmonize via denoising: Subject-driven image editing with pre-trained diffusion model. *arXiv:2306.07596*, 2023d.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *CVPR*, 2023e.

Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *CVPR*, 2024c.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, 2023f.

Zhongping Zhang, Jian Zheng, Zhiyuan Fang, and Bryan A Plummer. Text-to-image editing by image information removal. In *WACV*, 2024d.

Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *CVPR*, 2024a.

Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, and Wenjing Yang. Magicfusion: Boosting text-to-image generation performance by fusing diffusion models. In *ICCV*, 2023a.

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 *NeurIPS*, 2024b.

Yang Zhao, Yanwu Xu, Zhisheng Xiao, and Tingbo Hou. Mobilediffusion: Subsecond text-to-image generation on mobile devices. *arXiv:2311.16567*, 2023b.

Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *CVPR*, 2024a.

Qingping Zheng, Ling Zheng, Yuanfan Guo, Ying Li, Songcen Xu, Jiankang Deng, and Hang Xu. Self-adaptive reality-guided diffusion for artifact-free super-resolution. In *CVPR*, 2024b.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

Zhenyu Zhou, Defang Chen, Can Wang, and Chun Chen. Fast ode-based sampling for diffusion models in around 5 steps. In *CVPR*, 2024.

A Appendix

Table 2: The corresponding works for the various stacks of conditioning mechanisms shown in Tab.1.

Stack of conditioning mechanisms for denoising network	
Serial Number	**Model**
DN1	Rombach et al. (2022); Saharia et al. (2022b); Nichol et al. (2022) Ramesh et al. (2022); Gu et al. (2022); Balaji et al. (2022)
DN2	Saharia et al. (2022c); Ho et al. (2022); Li et al. (2022a)
DN3	Shang et al. (2024); Zhao et al. (2024a); Jiang et al. (2023a) Zheng et al. (2024a;b); Zhang et al. (2024c); Xue et al. (2024)
DN4	Fu et al. (2023); Huang et al. (2023c); Feng et al. (2023); Li et al. (2023e)
DN5	Brooks et al. (2023); Zhang et al. (2024b); Yildirim et al. (2023); Zhang et al. (2024a) Geng et al. (2023); Sheynin et al. (2024); Li et al. (2023f)
DN6	Kawar et al. (2023); Wu et al. (2023); Zhang et al. (2023c); Mou et al. (2024b) Mahajan et al. (2024); Ravi et al. (2023); Bodur et al. (2024); Li et al. (2023c); Zhang et al. (2023f)
DN7	Xiao et al. (2023); Ma et al. (2024); Shi et al. (2024a); Gal et al. (2023b) Jia et al. (2023); Lu et al. (2024); Shiohara & Yamasaki (2024)
DN8	Ruiz et al. (2023); Gal et al. (2023a); Kumari et al. (2023) Gu et al. (2024); Liu et al. (2023c;d); Han et al. (2023)
DN9	Mou et al. (2024c); Zhang et al. (2023b;d); Goel et al. (2023) Qin et al. (2023); Yang et al. (2023b); Zhao et al. (2024b); Ran et al. (2024); Jiang et al. (2023b)
DN10	Wang et al. (2022a); Xu et al. (2024)
DN11	Li et al. (2023h); Zeng et al. (2024); Kim et al. (2024)
DN12	Yang et al. (2023a); Xie et al. (2023a); Wang et al. (2023b); Xie et al. (2023b) Song et al. (2023c); Kim et al. (2023b); Chen et al. (2024c)
DN13	Li et al. (2023i); Hoe et al. (2023); Wang et al. (2024a)

Stack of conditioning mechanisms for sampling process	
Serial Number	**Model**
SP1	Tian et al. (2024); Liu et al. (2023b)
SP2	Luo et al. (2023); Yue et al. (2024); Welker et al. (2024); Wang et al. (2024c); Delbracio & Milanfar (2023)
SP3	Kawar et al. (2021; 2022); Wang et al. (2024b)
SP4	Avrahami et al. (2022); Chung et al. (2022a); Song et al. (2023a) Chung et al. (2023a); Fei et al. (2023); Rout et al. (2024)
SP5	Lugmayr et al. (2022); Choi et al. (2021); Chung et al. (2022b)
SP6	Su et al. (2023); Meng et al. (2022a); Mokady et al. (2023); Dong et al. (2023); Wang et al. (2023c); Wallace et al. (2023); Miyake et al. (2023); Ju et al. (2023); Meiri et al. (2023); Brack et al. (2024); Wu & De la Torre (2023); Huberman-Spiegelglas et al. (2023); Nie et al. (2023); Zhang et al. (2023e)
SP7	Couairon et al. (2023); Yang et al. (2023c); Patashnik et al. (2023) Wang et al. (2023a); Lee et al. (2024); Huang et al. (2023b)
SP8	Hertz et al. (2023); Tumanyan et al. (2023); Cao et al. (2023) Lu et al. (2023); Chung et al. (2024); Guo & Lin (2023)
SP9	Parmar et al. (2023); Mou et al. (2024b); Lin et al. (2023); Park et al. (2024)
SP10	Voynov et al. (2023); Singh et al. (2023); Luo et al. (2024); Mo et al. (2024)
SP11	Zhao et al. (2023a); Shirakawa & Uchida (2024); Bar-Tal et al. (2023)
SP12	Cao et al. (2023); Patashnik et al. (2023); Lu et al. (2023) Balaji et al. (2022); Liu et al. (2023d); Guo & Lin (2023)
SP13	Chen et al. (2024b); Epstein et al. (2024); Mou et al. (2024a)
SP14	Liu et al. (2022)
SP15	Ho & Salimans (2022); Sadat et al. (2024); Kynkäänniemi et al. (2024)
SP16	Yu et al. (2023); Bansal et al. (2023)