

# Controllable Generation with Text-to-Image Diffusion Models: A Survey

Pu Cao, Feng Zhou, Qing Song, Lu Yang

**Abstract**—In the rapidly advancing realm of visual generation, diffusion models have revolutionized the landscape, marking a significant shift in capabilities with their impressive text-guided generative functions. However, relying solely on text for conditioning these models does not fully cater to the varied and complex requirements of different applications and scenarios. Acknowledging this shortfall, a variety of studies aim to control pre-trained text-to-image (T2I) models to support novel conditions. In this survey, we undertake a thorough review of the literature on controllable generation with T2I diffusion models, covering both the theoretical foundations and practical advancements in this domain. Our review begins with a brief introduction to the basics of denoising diffusion probabilistic models (DDPMs) and widely used T2I diffusion models. We then reveal the controlling mechanisms of diffusion models, theoretically analyzing how novel conditions are introduced into the denoising process for conditional generation. Additionally, we offer a detailed overview of research in this area, organizing it into distinct categories from the condition perspective: generation with specific conditions, generation with multiple conditions, and universal controllable generation. For an exhaustive list of the controllable generation literature surveyed, please refer to our curated repository at <https://github.com/PRIV-Creation/Awesome-Controllable-T2I-Diffusion-Models>.

**Index Terms**—Survey, Text-to-Image Diffusion Model, Controllable Generation, AIGC

## 1 INTRODUCTION

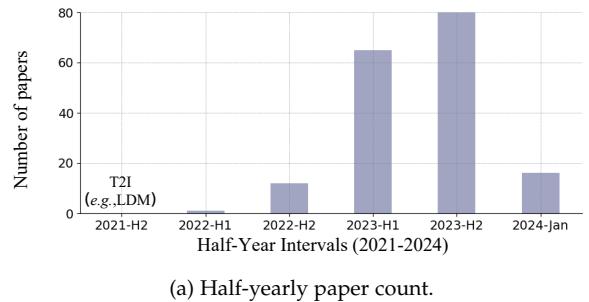
DIFFUSION models, representing a paradigm shift in the visual generation, have dramatically outperformed traditional frameworks like Generative Adversarial Networks (GANs) [1]–[8]. As parameterized Markov chains, diffusion models exhibit a remarkable ability to transform random noise into intricate images, progressing sequentially from noise to high-fidelity visual representations. With the advancement of technology, diffusion models have demonstrated immense potential in image generation and related downstream tasks.

As the quality of imagery generated by these models advances, a critical challenge becomes increasingly apparent: achieving precise control over these generative models to fulfill complex and diverse human needs. This task goes beyond simply enhancing image resolution or realism; it involves meticulously aligning the generated output with the user's specific and nuanced requirements as well as their creative aspirations. Fueled by the advent of extensive multi-modal text-image datasets [9]–[17] and development of guidance mechanism [18]–[21], text-to-image (T2I) diffusion models have emerged as a cornerstone in the controllable visual generation landscape [21]–[26]. These models are capable of generating realistic, high-quality images that accurately reflect the descriptions provided in natural language.

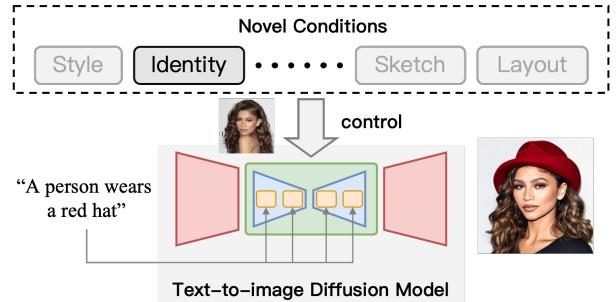
While text-based conditions have been instrumental in propelling the field of controllable generation forward, they inherently lack the capability to fully satisfy all user requirements. This limitation is particularly evident in scenarios where conditions, such as the depiction of an

• Pu Cao, Feng Zhou, Qing Song, Lu Yang are with the Beijing University of Posts and Telecommunications, Beijing, 100876, China (e-mail: caopu@bupt.edu.cn; zhousfeng@bupt.edu.cn; prior@bupt.edu.cn; soeaver@bupt.edu.cn)

• Corresponding author: Lu Yang (email: soeaver@bupt.edu.cn)



(a) Half-yearly paper count.



(b) Schematic diagram of controllable generation.

**Fig. 1: An overview of conditional generation with T2I diffusion model.** (a) We plot the number of papers on controllable generation based on T2I diffusion models, implying that it is increasing rapidly after powerful generators are released. (b) We present a schematic illustration of controllable generation using the T2I diffusion model, where novel conditions beyond text are introduced to steer the outcomes. Example images are sourced from [27].

unseen person or a distinct art style, are not effectively conveyable through text prompts alone. These scenarios pose significant challenges in the T2I generation process, as the nuances and complexities of such visual representations are difficult to encapsulate in text form. Recognizing this gap, a substantial body of research has shifted focus towards integrating novel conditions that extend beyond the confines of textual descriptions into T2I diffusion models. This pivot has been further facilitated by the emergence of powerful and open-sourced T2I diffusion models, as illustrated in Figure 1a. These advancements have led to the exploration of diverse conditions, thereby enriching the spectrum of possibilities for conditional generation and addressing the more intricate and nuanced demands of users in various applications.

There are numerous survey articles exploring the AI-generated content (AIGC) domain, including diffusion model theories and architectures [28], efficient diffusion models [29], multi-modal image synthesis and editing [30], visual diffusion model [31]–[34], and text-to-3D applications [35]. However, they often provide only a cursory brief of controlling text-to-image diffusion models or predominantly focus on alternative modalities. This lack of in-depth analysis of the integration and impact of novel conditions in T2I models highlights a critical area for future research and exploration.

This survey provides an exhaustive review of controllable generation using text-to-image diffusion models, encompassing both theoretical foundations and practical applications. Initially, we provide a concise overview of the background of T2I diffusion models and delve into the theoretical underpinnings of these methods, elucidating how novel conditions are integrated into T2I diffusion models. This exploration sheds light on the fundamentals of prior research and facilitates a deeper understanding of the field. Subsequently, we offer a thorough overview of previous studies, highlighting their unique contributions and distinguishing features. Additionally, we explore the varied applications of these methods, showcasing their practical utility and impact in diverse contexts and related tasks.

In summary, our contributions are:

- We introduce a well-structured taxonomy of controllable generation methods from the condition perspective, shedding light on the inherent challenges and complexities in this study area.
- We conduct an in-depth analysis of two core theoretical mechanisms essential for incorporating novel conditions into T2I diffusion models: conditional score prediction and condition-guided score estimation, providing a nuanced understanding of how these mechanisms function at a granular level.
- Our review is comprehensive, covering a wide range of conditional generation studies according to our proposed taxonomy. We meticulously underscore the salient features and distinctive characteristics of each method.
- We showcase the diverse applications of conditional generation using T2I diffusion models across various generative tasks, demonstrating its emergence as a fundamental and influential aspect in the AIGC era.

The rest of this paper is organized as follows. Section 2 provides a brief introduction to denoising diffusion probabilistic models (DDPMs), demonstrates the widely used text-

to-image diffusion models, and presents a well-structured taxonomy. In Section 3, we analyze the controlling mechanisms and reveal how to introduce novel conditions in text-to-image diffusion models. In Section 4, we summarize existing approaches for controlling the text-to-image diffusion model according to our proposed taxonomy. Finally, section 7 demonstrates the applications of controllable text-to-image generation.

## 2 PRELIMINARIES

### 2.1 Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Models (DDPMs) represent a novel class of generative models that operate on the principle of reverse diffusion. These models are formulated as parameterized Markov chains that synthesize images by gradually converting noise into structured data through a sequence of steps.

- **Forward Process.** The diffusion process begins with the data distribution  $x_0 \sim q(x_0)$  and adds gaussian noise incrementally over  $T$  timesteps. At each step  $t$ , the data  $x_t$  is noised by a transition kernel:

$$q(x_{1:T}|x_0) := \prod_{t=1}^T q(x_t|x_{t-1}), \quad (1)$$

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad (2)$$

where  $\beta_t$  are variance hyperparameters of the noise.

- **Reverse Process.** During the reverse process of a DDPM, the model's objective is to progressively denoise the data, thereby approximating the reverse of the Markov chain. This process begins from the noise vector  $x_T$  and transitions towards the original data distribution  $q(x_0)$ . The generative model parameterizes the reverse transition  $p_\theta(x_{t-1}|x_t)$  as a normal distribution:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (3)$$

where deep neural networks, often instantiated by architectures like UNet, parameterize the mean  $\mu_\theta(x_t, t)$  and variance  $\Sigma_\theta(x_t, t)$ . The UNet takes the noised data  $x_t$  and time step  $t$  as inputs and outputs the parameters of the normal distribution, thereby predicting the noise  $\epsilon_\theta$  that the model needs to reverse the diffusion process. To synthesize new data instances  $x_0$ , we initiate by sampling a noise vector  $x_T \sim p(x_T)$  and then successively sample from the learned transition kernels  $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$  until we reach  $t = 1$ , completing the reverse diffusion process.

### 2.2 Text-to-Image Diffusion Models

In this section, we spotlight several pivotal and widely utilized text-to-image foundational models. Detailed information regarding these models is systematically compiled and presented in Table 1.

- **GLIDE [21].** To generate images aligned with free-form text prompts, GLIDE intuitively replace the *class label* in class-conditioned diffusion models (*i.e.* ADM [18]) with *text*, formalizing the first text-to-image diffusion model. The authors explore two different guidance for text-conditioning.

For classifier guidance, GLIDE trains a CLIP model in noisy image space to provide *CLIP guidance*. Following [20], GLIDE additionally investigates classifier-free guidance (CFG) for comparison, which yields more preferable results in both image photo-realism and textual alignment by human evaluators and is chosen as the fundamental mechanism for text-to-image generation. For text condition, GLIDE first transforms the input text  $c$  into a token sequence via a trainable transformer [36]. Subsequently, they replace the class embedding with the pooled text features and further concatenate the projected sequence text features to the attention context at each attention layer in diffusion model. GLIDE trains the diffusion model and text transformer on the same dataset as DALL-E [22]. The diffusion model is trained to predict  $p(x_{t-1}|x_t, c)$  and generate images with CFG.

- **Imagen** [24]. Following GLIDE, Imagen adopts classifier-free guidance (CFG) for text-to-image generation. Instead of training a task-specified text encoder from scratch in GLIDE, Imagen leverages a pre-trained and frozen large language model (LLM) as its text encoder, aiming to reduce computational demands. The authors conduct a comparative analysis of various LLMs, including those trained on image-text datasets (*e.g.*, CLIP [39]) and solely on text corpora (*e.g.*, BERT [40], T5 [37]). Their findings suggest that increasing the scale of language models more effectively enhances the fidelity of samples and the congruence between image and text, compared to the enlargement of image diffusion models. Furthermore, Imagen’s exploration into different text conditioning methods reveals cross-attention as the most effective technique.

- **DALL-E 2** [38]. To leverage the robust semantic and style representations of images from contrastive models like CLIP [39], DALL-E 2, also known as unCLIP, trains a generative diffusion *decoder* to invert the CLIP image encoder. The generating process consists of the following steps. First, given an image caption  $y$  and its text embedding  $z_t$ , a *prior*  $p(z_i|z_t)$  bridges the gap between CLIP text and the image latent space, where  $z_i$  is the image embedding. Second, a *decoder*  $p(x|z_i)$  generates image  $x$  from the image embedding. Specifically, the *decoder* is a diffusion model modified from the architecture of GLIDE, where the CLIP embedding is projected and added to the existing time-step embedding. The *prior* can be optimized using either an autoregressive approach or a diffusion model, with the latter demonstrating superior performance.

- **Latent Diffusion Model (LDM)** [23]. To enable diffusion model training and inference on limited computational resources and generate high-resolution images of high quality and flexibility, LDM applies the denoising process in the latent space of pre-trained autoencoders. Specifically, the autoencoder  $\mathcal{E}$  maps images  $x \in \mathcal{D}_x$  into a spatial latent space  $z = \mathcal{E}(x)$ . To develop a conditional image generator, LDM enhances the underlying UNet with the cross-attention mechanism to effectively model the conditional distribution  $p(z_{t-1}|z_t, c)$ , where  $c$  is the conditional input, such as text prompts and segmentation masks.

In the realm of text-to-image generation, the authors employ the LAION-400M dataset to train a 1.45 billion parameter text-to-image LDM model, capable of producing

images at a resolution of  $256 \times 256$  (with a latent resolution of  $32 \times 32$ ). For the encoding of text inputs, a BERT tokenizer [40] was utilized as the text encoder.

- **Stable Diffusion (SD).** Built upon the Latent Diffusion Model (LDM) framework, Stability AI developed and launched several series of text-to-image diffusion models, termed Stable Diffusion. SD demonstrates unparalleled capabilities in text-to-image generation, and with its models being open-sourced, it has gained widespread usage within the community.

### 2.3 Taxonomy

The task of conditional generation utilizing text-to-diffusion models represents a multifaceted and intricate domain. From the condition perspective, we divide this task into three sub-tasks (refer to Figure 2). Most works study how to generate images under specific conditions, *e.g.* image-guided generation, and sketch-to-image generation. To reveal the mechanical theory and features of these approaches, we further categorize them according to their condition types. The primary challenge in this task lies in how to enable pretrained text-to-image (T2I) diffusion models to learn to model new types of conditions and to generate in conjunction with textual conditions while ensuring the images produced are of high quality. Additionally, some methods investigate how to generate images using multiple conditions, such as given a character’s identity and pose. The main challenge in these tasks is the integration of multiple conditions, necessitating the capability to express several conditions simultaneously in the generated results. Furthermore, some works attempt to develop a condition-agnostic generation approach that can utilize these conditions to produce results.

## 3 HOW TO CONTROL TEXT-TO-IMAGE DIFFUSION MODELS WITH NOVEL CONDITIONS

In this section, we present the controlling mechanism of diffusion models from a score-based perspective [19]. Following [188], we can set the approximate denoising transition mean  $\mu_\theta(x_t, t)$  in Equation.3 as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}} s_\theta(x_t, t) \quad (4)$$

where  $s_\theta(x_t, t)$  is a neural network that learns to predict the score function  $\nabla_{x_t} \log p_t(x)$ . In DDPM, we have:

$$\nabla_{x_t} \log p_t(x) = -\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \epsilon \quad (5)$$

where  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$  is the gaussian noise used in forward process,  $\alpha_t := 1 - \beta_t$ , and  $\bar{\alpha}_t := \prod_{s=0}^t \alpha_s$ . Then, Equation.4 can be written as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\epsilon}(x_t, t) \right) \quad (6)$$

where  $\hat{\epsilon}(x_t, t)$  predicts  $\epsilon$ .

In conditional generation ( $c$  denotes condition), the score function is extended with a posterior probability term  $\nabla_{x_t} \log p_t(c|x)$  and becomes  $\nabla_{x_t} \log (p_t(x)p_t^w(x|c))$  ( $w$  represents a hyper-parameter to control condition intensity), following [18], [20]. To employ a neural network for

TABLE 1: Collection of primary and used text-to-image diffusion models in this survey.  $\dagger$ : number of UNet and text encoder's parameters (default refers only to UNet).  $f$ : downsampling factor of autoencoder in latent-space diffusion models. CLIP: open source implementation of CLIP. \*: train from scratch.

Model	Publication	Param.	Resolution	$f$	Text Encoder	Training Dataset	Open Source
<b>Pixel Space Diffusion Models</b>							
GLIDE [21]	ICML 2022	5.0B $\dagger$	$256^2$	-	plain Transformer* [36]	DALL-E [22]	✓
Imagen [24]	NeurIPS 2022	3.0B	$1024^2$	-	T5-XXL [37]	>LAION-400M [16]	✗
DALL-E 2 [38]	arXiv 2022	4.5B	$1024^2$	-	CLIP* [39] & Diffusion prior*	CLIP [39] & DALL-E [22]	✗
<b>Latent Space Diffusion Models</b>							
LDM [23]	CVPR 2022	903M	$256^2$	8	BERT-tokenizer [40]	LAION-400M [16]	✓
SD v1.x [23]	CVPR 2022	860M	$512^2$	8	CLIP-ViT-L/14 [39]	LAION-2B [17]	✓
SD v2.x [23]	CVPR 2022	865M	$512^2/768^2$	8	CLIP-ViT-H/14 [39]	LAION-5B [17]	✓
SD XL [25]	ICLR 2024	2.6B	$1024^2$	8	CLIP-ViT/G & CLIP-ViT/L [39]	internal dataset	✓

conditional generation, classifier-free guidance (CFG) [20] transforms it to:

$$\begin{aligned} \nabla_{x_t} \log(p_t(x)p_t^w(x|c)) \\ = \nabla_{x_t} \log p_t(x) + w \nabla_{x_t} \log p_t(c|x) \\ = \nabla_{x_t} \log p_t(x) + w \nabla_{x_t} \log \frac{p_t(x|c)}{p_t(x)} \\ = (1-w) \nabla_{x_t} \log p_t(x) + w \nabla_{x_t} \log p_t(x|c) \end{aligned} \quad (7)$$

where  $\nabla_{x_t} \log p_t(x)$  and  $\nabla_{x_t} \log p_t(x|c)$  can be predicted by training a model  $\epsilon_\theta(x_t, \cdot, t)$ , which predict the former via  $\epsilon_\theta(x_t, \phi, t)$  and the latter via  $\epsilon_\theta(x_t, c, t)$ . Existing T2I diffusion models train  $\epsilon_\theta(x_t, \cdot, t)$  by randomly dropping the text prompt, and the denoising process with CFG is as follows:

$$\hat{\epsilon}(x_t, c_{text}, t) = (1-w)\epsilon_\theta(x_t, \phi, t) + w\epsilon_\theta(x_t, c_{text}, t) \quad (8)$$

and  $\hat{\epsilon}(x_t, c_{text}, t)$  is used in Equation 4 for conditional synthesis.

Hence, the key to controlling text-to-image models with novel conditions  $c_{novel}$  is to model score  $\nabla_{x_t} \log p_t(x|c_{text}, c_{novel})$ . Following [18], [32], there are two types of mechanisms, i.e., conditional score prediction and conditioned-guided score estimation, which we illustrate below.

### 3.1 Conditional Score Prediction

While T2I diffusion models leverage  $\epsilon_\theta(x_t, c_{text}, t)$  to predict  $\nabla_{x_t} \log p_t(x|c_{text})$ , a fundamental and powerful way for steering diffusion models is through conditional score prediction in the sampling process, where these methods introduce  $c_{novel}$  into  $\epsilon_\theta(x_t, c_{text}, t)$ , constructing a  $\tilde{\epsilon}(x_t, c_{text}, c_{novel}, t)$  to straightforwardly predict  $\nabla_{x_t} \log p_t(x|c_{text}, c_{novel})$ . Then, the denoising process with CFG of conditional score prediction methods is as follows:

$$\begin{aligned} \hat{\epsilon}(x_t, c_{text}, c_{cond}, t) = (1-w)\tilde{\epsilon}(x_t, \phi, t) \\ + w\tilde{\epsilon}(x_t, c_{text}, c_{cond}, t) \end{aligned} \quad (9)$$

We here illustrate several mainstream ways to attain  $\tilde{\epsilon}(x_t, c_{text}, c_{novel}, t)$ .

- **Model-based Conditional Score Prediction.** Some approaches employs an additional encoder  $E$  to encode novel

conditions and input the encoded features into  $\epsilon_\theta$ , where the conditional score prediction process is as follows:

$$\tilde{\epsilon}(x_t, c_{text}, c_{novel}, t) = \epsilon_{\theta^*}(x_t, c_{text}, E(c_{novel}), t) \quad (10)$$

where  $E$  and  $\theta^*$  are trainable. The schematic illustration is shown in Figure 3a.

- **Tuning-based Conditional Score Prediction.** Tuning-based methods typically focus on adapting to a specific condition, often in scenarios with limited data, such as single or few-shot examples. These methods achieve conditional prediction by transforming either the text condition  $c_{text}$  or the model parameters  $\theta$  into a form specific to the given condition, as shown in Figure 3b. This can be represented as:

$$\tilde{\epsilon}(x_t, c_{text}, c_{novel}, t) = \epsilon_{\theta^*}(x_t, c_{text}^*, t) \quad (11)$$

where condition information is memorized in  $c_{text}^*$  and  $\theta$ .

- **Training-free Conditional Score Prediction.** While the above techniques require a training process, some methods are designed in a training-free manner (refer to Figure 3c). They introduce conditions to control the generation directly through the intrinsic ability of the structure of UNet, such as modulating the cross-attention map to control the layout [135], [142] or introducing features of the reference image in self-attention to control the style [101].

### 3.2 Condition-Guided Score Estimation

Unlike conditional score prediction approaches predicting  $\nabla_{x_t} \log p_t(x|c_{text}, c_{novel})$ , condition-guided estimation approaches are designed to gain  $\nabla_{x_t} \log p_t(c_{novel}|x_t)$  without need of CFG, which generally train an additional model with parameters  $\varphi$  to predict the condition from latent or internal features, denoting as  $p_\varphi(c_{novel}|x_t)$ . It can be utilized to attain  $\nabla_{x_t} \log p_t(c_{novel}|x)$  via backpropagation, as illustrated in Figure 4. And the denoising process now reads:

$$\begin{aligned} \hat{\epsilon}(x_t, c_{text}, c_{novel}, t) = & \hat{\epsilon}(x_t, c_{text}, t) \\ & + \gamma \nabla_{x_t} \log p_\varphi(c_{novel}|x_t) \end{aligned} \quad (12)$$

where  $\gamma$  is a hyper-parameter to adjust the conditional score and  $\hat{\epsilon}(x_t, c_{text}, t)$  is the original score prediction of text-conditioned diffusion models with CFG.

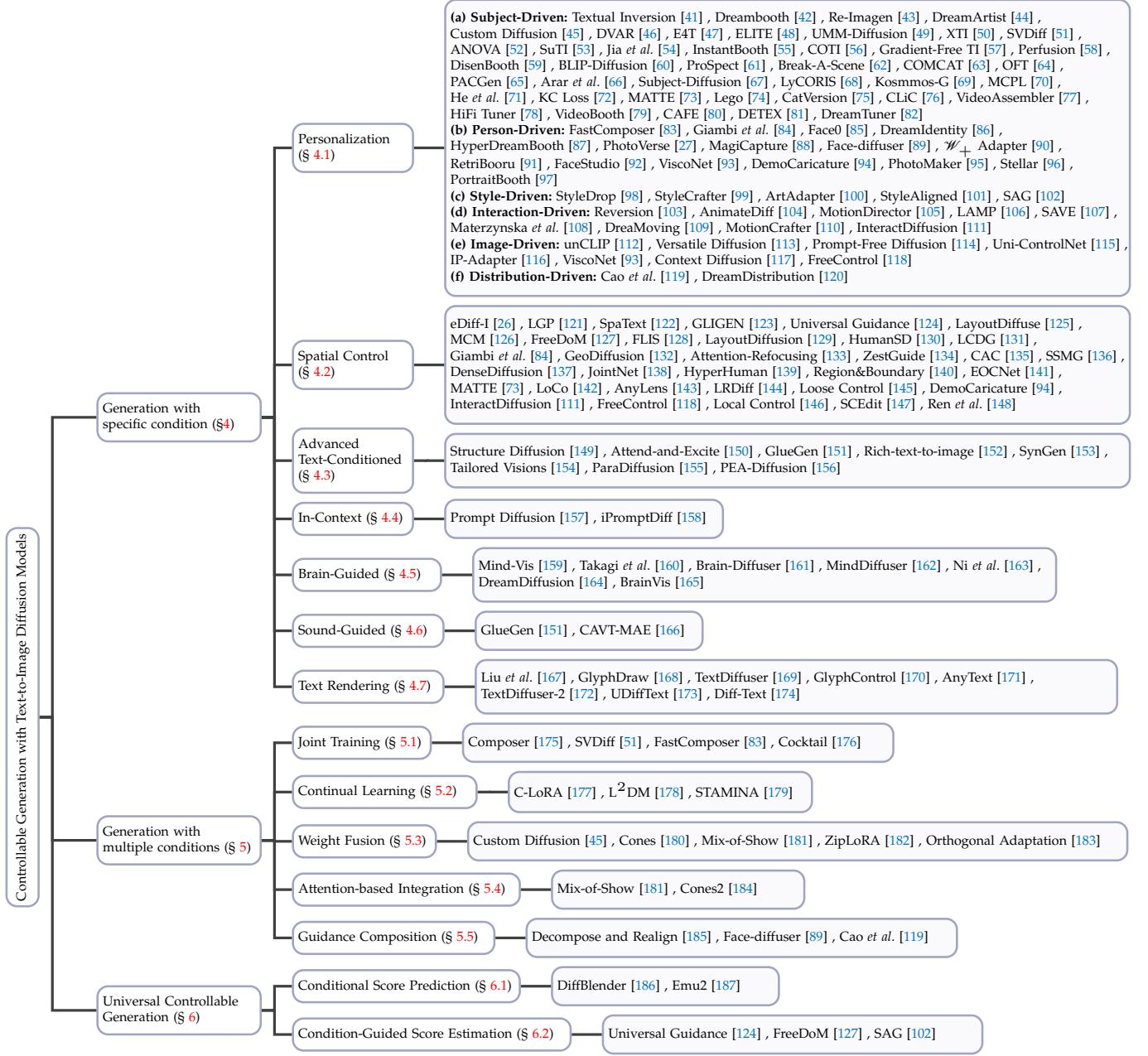


Fig. 2: **Taxonomy of Controllable Generation.** From the condition perspective, we categorize controllable generation approaches into three sub-tasks, including generation with specific conditions, generation with multiple conditions, and universal controllable generation.

## 4 CONTROLLABLE TEXT-TO-IMAGE GENERATION WITH SPECIFIC CONDITIONS

Building upon the foundation of text-to-image diffusion models, introducing novel conditions to steer the generative process represents a complex and multifaceted task. In the following chapters, we review the existing methods of conditional generation according to the condition perspective, providing a comprehensive critique of their methodologies.

### 4.1 Personalization

Personalization task aims to capture and utilize concepts as generative conditions, which are not easily describable

through text, from exemplar images for controllable generation. In this section, we provide an overview of these personalized conditions, categorizing them to offer a clearer understanding of their diverse applications and functionalities. We illustrate the results of personalization in Figure 5.

#### 4.1.1 Subject-Driven Generation

In this section, we provide a detailed overview of subject-driven generation methods. The subject-driven generation task (also known as subject-centric personalization) is designed to produce visual content that retains the subjects of provided samples. In practice, many subject-driven generation methods are not confined to conditions specific

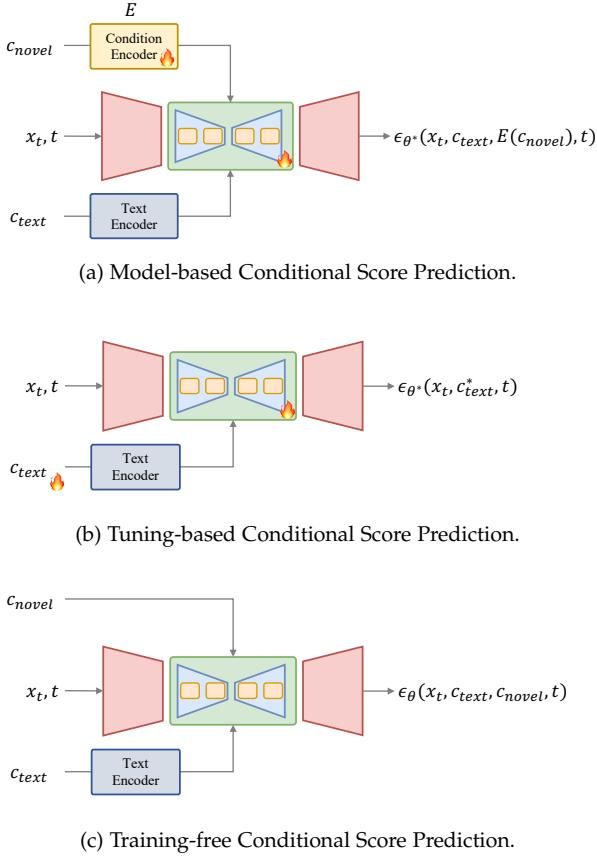


Fig. 3: Illustrations of conditional score prediction mechanisms. .

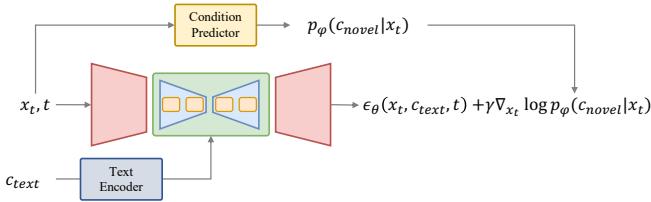


Fig. 4: Illustration of condition-guided score estimation.

to subject types; they often demonstrate a more universal capability. Thus, many of the approaches discussed in this chapter can be extended to a wider range of customized tasks. In summarizing these works, we adopt a broader perspective to showcase their general applicability as much as possible, aiming to facilitate a better understanding of their contributions and roles.

According to the controlling mechanism mentioned in Section 3, since all of these methods use conditional score prediction to introduce the conditions, we category them by their pipelines: tuning-based methods, which adapt model parameters or embeddings to cater to specific conditions; model-based methods, employing encoders to extract personalized conditions and feeding them into diffusion models; and training-free methods, which leverage external references to steer the generative process without the need of training.

- **Tuning-based Personalized Score Prediction.** A simple

yet effective way to grasp concepts from provided samples involves selectively tuning a subset of parameters to reconstruct these concepts within text-to-image models, where the updated parameters are tailored to the desired concepts [41], [42], [44], [50], [61], [72], [77].

As the basic input for text-to-image diffusion models, text plays a crucial role in adapting these models to specific user needs. Textual Inversion (TI) [41] adopts an innovative approach by embedding user-provided concepts into new ‘words’ within the text embedding space. This method expands the tokenizer’s dictionary and optimizes additional tokens using a denoising process on provided images. DreamBooth [42] follows a similar path but utilizes low-frequency words (*i.e.*, sks) to represent concepts and additionally updates the parameters of the UNet with a class-specific prior preservation loss to enhance the diversity of generated outputs. The straightforward and adaptable frameworks of TI and DreamBooth establish them as foundational models for numerous subsequent tuning-based methods. Furthermore, Custom Diffusion [45] analyzes weight deviations during the fine-tuning process and discovers the pivotal role of cross-attention layer parameters, particularly key and value projections (*i.e.*,  $W^k$  and  $W^v$ ). This insight leads to a focused update on these projections and the incorporation of extra text tokens and regularization loss for fine-tuning.

Some approaches have been taken to expand the text embedding space, particularly by considering the distinction of each UNet layer [50], [61]. They apply distinct text embeddings across various layers. In contrast, CatVersion [75] diverges from the focus on text embeddings and the UNet’s parameters and advocates for tuning concatenated embeddings within the feature-dense space of the text encoder. Such a method is suggested to be more effective in learning the nuances between a personalized concept and its base class, contributing to the preservation of prior knowledge within the model.

In addition, **parameter-efficient tuning (PEFT)** [191]–[194] plays a pivotal role in personalization methods [63]. Low-rank Adaptation (LoRA) [192] has seen widespread integration across a range of personalization techniques [42], [59], [104], [177], [181]. Besides, Xiang *et al.* propose ANOVA [52], which opts for the adapter [191], and reveals that placing adapters subsequent to the cross-attention block enhances performance significantly. To facilitate the comprehensive application and evaluation of PEFT in the fine-tuning of diffusion models, LyCORIS [68] has developed an open-source library<sup>1</sup>. This library encompasses a broad spectrum of PEFT methods, including but not limited to LoRA [192], LoHa, and DyLoRA [193]. LyCORIS further introduces a detailed framework for the systematic analysis and assessment of these PEFT techniques, significantly advancing the field of diffusion model personalization.

Moreover, a critical challenge in the realm of personalization is the **disentanglement** of specific concepts from the provided samples. Numerous studies [59], [62], [65], [74], [81] have identified a common issue where extraneous information becomes intertwined with the intended concept during the customization process, such as inadvertently learning the context surrounding images in subject-driven

1. <https://github.com/KohakuBlueleaf/LyCORIS>

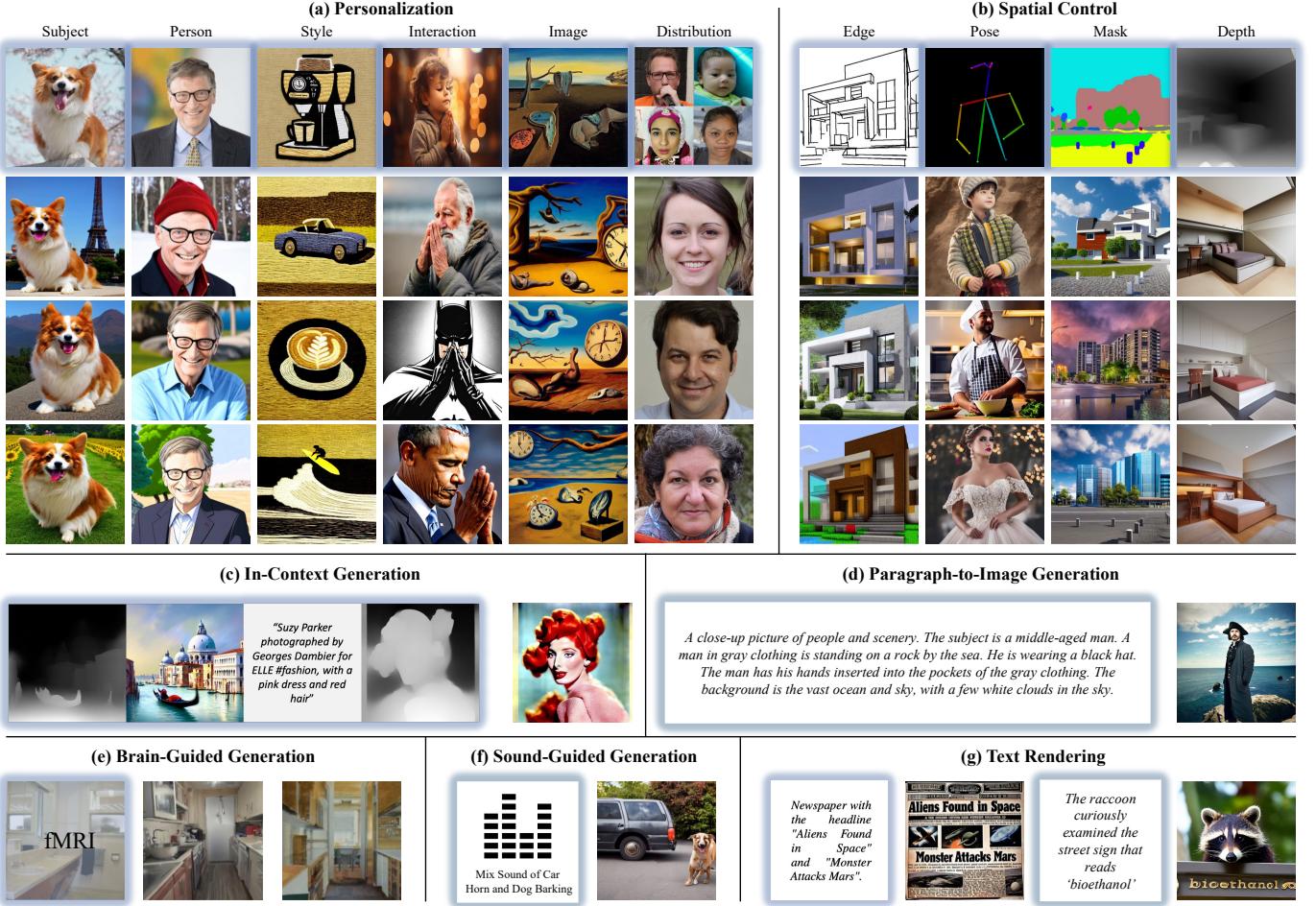


Fig. 5: Illustration of controllable text-to-image generation with specific conditions. The condition is marked in blue background. Examples are sourced from [27], [48], [100], [112], [119], [151], [155], [157], [162], [174], [189], [190].

generation. To effectively isolate and extract the essential concept information from samples, several works [62], [70], [76] have investigated the use of explicit masks. In a similar vein, Disenbooth [59] and DETEX [81] focus on mitigating the influence of background elements in the personalization process. DETEX goes a step further by also aiming to decouple the pose information of subjects from the overall concept. Meanwhile, PACGen [65] employs aggressive data augmentation techniques, transforming the size and location of the personalized concepts within the samples, thereby aiding in the separation of spatial information from the core concept itself.

Also, training diffusion models on small-scale datasets often meet another significant challenge: it risks compromising the generative model's broader applicability, leading to the requirement of a delicate balance between fidelity and editability [82]. To tackle this issue, several studies have introduced **preservation mechanisms**, focusing on strategies to prevent overfitting to input samples [42], [45], [51], [58], [64], [78]. For instance, Perfusion [58] addresses this by locking concepts' cross-attention keys to their prior categories and employing a gated rank-1 method for concept learning. SVDiff [51] takes a different approach, adjusting the singular values in the weight matrices of the model.

This technique is designed to minimize overfitting risks and mitigate issues like language drifting. Furthermore, OFT [64] emphasizes the importance of the hyperspherical energy in weight matrices for sustaining the model's semantic generative capabilities. Hence, it introduces an orthogonal fine-tuning method, further contributing to the preservation of the model's generalization ability in the face of limited training data.

In addition to the aforementioned methods, several researchers have explored alternative training techniques aimed at optimizing generative performance, expediting the tuning process, and minimizing GPU memory usage [46], [56], [57], [71], [73], [75]. Specifically, DVAR [46] identifies limitations in standard training metrics for assessing the convergence in concept learning and utilize a simple variance-based early stopping criterion, enhancing the efficiency of the fine-tuning process. Gradient-Free Textual Inversion [57] adopts an innovative approach by dividing the optimization process into two parts: dimension reduction in the search space and non-convex, gradient-free optimization in a subspace. This method achieves a significant speed-up in optimization with minimal impact on performance. MATTE [73] delves into the roles of timesteps and UNet's layers in personalizing various concept categories like color,

object, layout, and style, aiming to enhance performance across different concept types. Addressing the need for high-quality data in Textual Inversion, COTI [56] introduces an active and controllable data selection framework to improve Textual Inversion by broadening the data scope. Similarly, He *et al.* [71] adopt a data-centric approach, proposing a novel strategy for generating regularization datasets at both text and image levels, further enriching the research landscape in this domain.

- **Model-based Personalized Score Prediction.** Model-based methods employ encoders to embed concepts, offering a significant speed advantage over tuning-based approaches when extracting concepts from images. Some works focus on domain-aware encoders specifically designed to embed images from targeted domains [47], [55]. For instance, InstantBooth [55] employs a specialized encoder and adapters trained on the face and cat domains to extract text embeddings and detailed patch features for concept learning. In contrast, other model-based methods opt for a domain-agnostic approach, training encoders on open-world images to extract more generalized conditions [48], [49], [53], [66], [67], [69], [77], [79]. These methods typically utilize large pre-trained models like CLIP [39] and BLIP-2 [195] as image encoders, focusing on fine-tuning a limited number of parameters, such as a projection layer [48], [49], [66], [82]. ELITE [48], for example, integrates a global mapping network and a local mapping network based on CLIP [39]. The global network transforms hierarchical image features into multiple text embeddings, while the local network infuses patch features into cross-attention layers for detailed reconstruction. BLIP-Diffusion [60] advances customization by pre-training a BLIP-2 [195] encoder for text-aligned image representation and developing a task for learning subject representations, enabling the generation of novel subject renditions. Following on E4T [47], Arar *et al.* [66] introduce an encoder for acquiring text embeddings and propose a hypernetwork to predict LoRA-style attention weight offsets in UNet. SuTI [53] takes a unique approach inspired by apprenticeship learning [196], training a vast array of expert models on millions of internet image clusters. The apprentice model is then taught to imitate these experts' behaviors. CAFE [80] build a customization assistant based on pre-trained large language model and diffusion model.

- **Training-free Personalized Score Prediction.** The pivotal technique for training-free personalization is extracting concept information from reference images in synthesis process. Similar to retrieval-augmented generation in natural language processing, leveraging knowledge from samples helps models faithfully generate given concepts. Re-Imagen [43] represents a novel approach to generating images of uncommon or rare categories, such as Chortai(dog) and Picarones(food). This method leverages an external multimodal knowledge base, utilizing relevant image-text pairs retrieved from this database as references for image generation. Beyond this, several methods, whether tuning-based or model-based, incorporate the use of reference images to enhance the accuracy and fidelity of the visual details in the generated images [77], [78], [91].

#### 4.1.2 Person-Driven Generation

The person-driven generation task (also known as human-centric personalization) is specifically focused on creating human-centric visual outputs that maintain the same identity as the individuals depicted in the exemplar samples. While person-driven generation is a specialized subset of the broader subject-driven generation category and several methods pertinent to this task have already been discussed in the previous section, we will concentrate on highlighting and analyzing those techniques that are explicitly tailored for person-driven generation in this part.

Similar to model-based subject-driven generation, many person-driven approaches encode facial images into text embedding space to provide identity condition [27], [83], [85]–[87]. For example, to achieve a balance between identity preservation and editability, Xiao *et al.* [83] introduces a novel approach that combines text prompts with visual features derived from reference images of individuals, called Fast-Composer. Specifically, this method fuses the human-related text embeddings (e.g., 'man' and 'woman') with visual features by multilayer perceptron, effectively encapsulating both the textual and visual conditions of the person's identity. Besides CLIP [39], Face0 [85] and DreamIdentity [86] employs pre-trained face recognition models [197] as their facial image encoders, where Face0 utilizes the Inception ResNet V1 [198] and DreamIdentity introduces a ViT-style [199] Multi-word Multi-scale (M<sup>2</sup>ID) encoder. While most methods utilize multi-modal pre-trained image encoders (e.g., CLIP [39]) or facial recognition models [197],  $\mathcal{W}$  + Adapter [90] introduces an innovative approach using StyleGAN's inversion encoder.

Inspired by retrieval-augmented generation (RAG), Tang *et al.* [91] introduce a novel retrieval-based method specifically tailored for human-centric personalization. Complementing this approach, they also present an anime figures dataset, named RetriBooru-V1, which is uniquely characterized by enhanced identity and clothing labels. Central to their method is the use of a frozen Variational Autoencoder (VAE) [200] for encoding reference images and seamless integration into the generation process by cross-attention and zero-convolution layers. These layers play a crucial role in accurately positioning the reference attributes—such as identity and clothing features—at the correct geometric locations in the generated image, thereby ensuring a high degree of fidelity and relevance in the output.

In contrast to subject-driven methods, person-driven generation approaches can benefit significantly from face segmentation, obtained either through parsing models or annotations [27], [83], [95]–[97]. For instance, some works, such as Stellar [96], employ face masks to eliminate background elements during data processing, thereby sharpening the focus on human identity within the input data. Conversely, other approaches leverage face masks for constructing [27], [83], [95], [97] or adjusting [88] loss functions.

#### 4.1.3 Style-Driven Generation

The style-conditioned generation task aims to extract style information from given samples as conditions for controllable generation.

Similar to the approach of tuning-based subject-driven methods, StyleDrop [98] employs the fine-tuning of

adapters [191], [201] on Muse [198] to tailor the model to specific style conditions and further proposes a prompt engineering technique to construct training data that effectively separates subject cues from style. Concurrently, several methods are exploring the incorporation of encoders designed to generate style-related embeddings for conditional generation [99], [100]. In addressing the issue of content borrowing from style references, ArtAdapter [100] introduces an innovative Auxiliary Content Adapter (ACA), which is designed to furnish the UNet with essential content cues, thereby ensuring that the model maintains a focus on style elements.

Besides, several approaches attempt to establish training-free frameworks for style-consistent image generation [101], [102]. For instance, StyleAligned [101] is designed to produce a series of images that adhere to a given reference style. This method introduces a novel attention sharing mechanism within the self-attention layers, which facilitates the interaction between the features of individual images and those of an additional reference image. Such a design enables the generation process to consider and incorporate style elements from multiple images simultaneously. Additionally, StyleAligned enhances the alignment of style attributes by normalizing both queries and keys using the Adaptive Instance Normalization (AdaIN) [202], further refining the style consistency across the generated images.

#### 4.1.4 Interaction-Driven Generation

The interaction-conditioned generation task is specifically designed to learn and generate interaction-related concepts, such as human actions and human-object interactions (HOI). Essentially, this task centers around the novel idea of using a “verb” as the conditioning element.

For action-driven image generation, Huang *et al.* [189] proposes an Action-Disentangled Identifier (ADI) to decouple subject identity and action for improved action condition learning. To block the inversion of action-agnostic features, ADI extracts the gradient invariance from the constructed sample triples and masks the updates of irrelevant channels, which effectively ensures that the action condition is embedded in text embedding.

Moreover, Reversion [103] has been developed to comprehend relationships depicted in sample images, such as in scenarios where ‘object A <is painted on> object B,’ with <is painted on> serving as the personalized condition. The method introduces a novel relation-steering contrastive learning mechanism, uniquely utilizing prepositions as positive samples to accurately guide the relational prompt, while other words are treated as negative samples. Additionally, Reversion employs a relation-focal importance sampling technique, which prioritizes the selection of samples with higher levels of noise during training, which facilitates the model’s learning of high-level semantic relationships.

Tian *et al.* [111] introduce the InteractDiffusion model to encapsulate human-object interaction (HOI) information for controllable generation. Central to their methodology is the construction of triplet labels encompassing a person, an action, and an object, along with corresponding bounding boxes, which are tokenized by interaction embeddings in InteractDiffusion to learn and represent the intricate relationships between these subjects.

#### 4.1.5 Image-Driven Generation

The image-conditioned generation task aims to generate a similar image from multiple perspectives (*e.g.*, content and style) by using an exemplar image as the prompt.

unCLIP [38] is an early work to explore using image prompt for image generation, proposing a two-stage model: a prior that generates a CLIP image embedding given a text caption, and a decoder that generates an image conditioned on the image embedding. Xu *et al.* [113] expand the existing single-flow diffusion pipeline into a multi-task multimodal network, dubbed Versatile Diffusion (VD), that handles multiple flows of text-to-image, image-to-text, and variations in one unified model. Xing *et al.* [114] introduce the Prompt-Free Diffusion to discard text with image in a text-to-image model, only using visual inputs to generate new images. They propose a Semantic Context Encoder (SeeCoder), consisting of a backbone encoder, a decoder, and a query transformer [195], to encode exemplar image. During inference, the SeeCoder will replace the CLIP text encoder in Stable Diffusion to take the reference image as input. IP-Adapter [116] decouples cross-attention mechanism that separates cross-attention layers for text features and image features to achieve image prompt capability for the pretrained text-to-image diffusion models. While these methods straightforwardly use images as prompts, ViscoNet [93] use segment person images to provide fashion reference for human-centric generation.

#### 4.1.6 Distribution-Driven Generation

The distribution-conditioned generation task is designed to understand and learn from the data distribution of multiple exemplar images, with the aim of generating a variety of results reflective of this distribution. This approach is distinct from subject-centric personalization, as it focuses on adapting text-to-image models to generate broader, more abstract concepts or categories rather than individual subjects.

Cao *et al.* [119] introduce the Guidance-Decoupled Personalization framework, designed for generating specific concepts (*e.g.*, faces) with a high degree of fidelity and editability. This framework uniquely decouples the conditional guidance into two distinct components: concept guidance and control guidance. The concept guidance component is specifically trained to steer the sampling process in a manner that adheres to the underlying data distribution, thereby ensuring the accurate generation of the reference concept. Moreover, DreamDistribution [120] is proposed for learning a prompt distribution, keeping a set of learnable text embeddings to model their distribution at CLIP text encoder feature space. Then a reparameterization trick is utilized to sample from this distribution and update the learnable embeddings.

### 4.2 Spatial Control

Since text is challenging to represent structure information, *i.e.* position and dense label, controlling text-to-image diffusion methods with spatial signals, *e.g.* layout [10], [203], human pose [204]–[207], human parsing [208]–[212], and segmentation mask [213], [214], is a significant research field in diffusion models. Within this context, we commence with a brief overview of some unified methods for spatial

control, followed by a more detailed exploration into various specific categories of structure, such as bounding boxes and keypoints.

#### 4.2.1 Spatial-Conditional Score Prediction

In the domain of spatial-conditional score prediction, methods are developed to model  $\hat{e}_\theta(x_t, c_{text}, c_{spatial}, t)$  with the aim of generating results aligned with a given spatial condition  $c_{spatial}$ . We here overview model-based and training-free spatial-conditioned score prediction methods, since tuning-based methods [73] conceptualize structural conditions (*e.g.*, layout) more abstractly and do not explicitly utilize spatial conditions.

- **Model-based Score Prediction.** ControlNet [190] stands out among generalized spatial controlling methods, gaining recognition as a seminal work and winning the prestigious Marr Award in 2023. Different significantly from methods that simply tune the parameters of the original diffusion model [45], [87], ControlNet introduces an innovative architecture by incorporating an additional encoder copy within the UNet structure. This added encoder is connected with the original UNet layers through the proposed “zero convolutions” to prevent overfitting and catastrophic forgetting. The simplicity and adaptability of ControlNet’s architecture have not only proven effective but also led to its widespread adoption as a baseline in numerous subsequent studies [82], [136], [176], [186], [215]–[218]. Similarly, T2I-Adapter [219] is proposed to align internal knowledge in text-to-image diffusion models and external control signals. SCEdit [147] propose an efficient generative tuning framework, which integrates and edits skip connections using a lightweight tuning module named SC-Tuner.

While ControlNet [190] necessitates training distinct models for each type of controlling signal, some researchers have pursued the development of more generalized methods capable of handling a variety of spatial signals [115], [176], [215], [216]. To address this challenge, Qin *et al.* [215] introduce a task-aware HyperNet designed to modulate diffusion models for adaptability to different types of conditions, named UniControl. In this approach, conditions are encoded using a mixture of experts (MOE) adapter. Simultaneously, task instructions are transformed into task embeddings through the task-aware HyperNet, which are integrated to zero convolution for precise modulation of the condition features’ injection into the model.

In the arena of layout-conditioned score prediction, various innovative approaches have been introduced [111], [122], [123], [126], [128], [129], [132], [136], [141], [143], [144]. GLIGEN [123] employs grounded language as the basis for generation, embedding this grounding information into new trainable layers through a gated mechanism, thus enabling more controlled generation. In addition, SpaText [122] constructs spatio-textual representation by introducing CLIP image embeddings, where they stack these object embeddings in the same shapes and positions of the segments to control layout. Besides, some works focus on the face domain, synthesizing face images under face parsing condition [84], [125], [148], [148].

Several approaches have been developed to jointly denoise the spatial structure conditions for enhanced spatial

control. JointNet [138], an extension of a pre-trained text-to-image diffusion model, introduces a new branch for dense modalities such as depth maps, where a duplicate of the original network is intricately connected with the RGB branch, facilitating complex interactions between different modalities. Additionally, Liu *et al.* [139] propose the Latent Structural Diffusion Model, which innovatively denoises depth and surface normal along with the RGB image synthesis.

While the above methods hope to generate images fully aligned to given conditions, some methods study employing coarse and incomplete spatial condition [115], [145], [146]. Specifically, LooseControl [145] extracts proxy depth for 3D box control from images and finetunes the ControlNet [190] by LoRA [192], enabling to create complex environments (*e.g.*, rooms, street views) by specifying only scene boundaries and locations of primary objects.

- **Training-free Score Prediction.** Since attention mechanism explicitly model the relationships between text and image tokens, modulating attention map becomes a pivotal training-free technique for controlling structure in score prediction [26], [118], [135], [137], [142]. eDiff-I [26] presents a technique named “*paint-with-words*” (also known as pww), rectifying the cross-attention maps of each word by the correspondence segmentation maps to control the location of objects. Additionally, DenseDiffusion [137] introduces a more extensive modulation method by devising multiple regularization, enhancing the precision and flexibility of layout control in score prediction.

#### 4.2.2 Spatial-Guided Score Estimation

While numerous methods adhere to the paradigm of conditional score prediction like ControlNet, some studies have explored spatial controlling through spatial-guided score estimation [121], [124], [127], [131]. Notably, LGP [121] stands as an early pioneer, which innovatively introduces a Latent Edge Predictor, designed to extrapolate sketch information from a series of intermediate features within a UNet architecture. It employs the degree of similarity between condition sketch and predicted sketch to compute gradients, which are then utilized to guide the score estimation process. Its methodologies and insights have been a source of inspiration for numerous subsequent research endeavors in this field [131], [133], [134], [220]. ZestGuide [134] leverages segmentation maps extracted from cross-attention layers, aligning generation with input masks through gradient-based guidance during denoising.

### 4.3 Advanced Text-Conditioned Generation

While text serves as the fundamental condition in text-to-image diffusion models, several challenges persist in this domain. First, text-guided synthesis, particularly with complex text involving multiple subjects or enriched descriptions, often encounters issues of textual misalignment. Furthermore, the predominant training of these models on English datasets has led to a notable lack in multilingual generation capabilities. To address this limitation, innovative approaches aimed at expanding the linguistic scope of these models have been proposed.

- **Improving Textual Alignment.** Textual alignment plays a pivotal role in text-to-image diffusion models, providing

essential control over the generation process. Despite being trained on multimodal text-image datasets, these generative models often struggle to precisely capture and reflect the full spectrum of information contained in textual descriptions. To address this challenge, various innovative approaches have been developed [149], [150], [152]. Specifically, Attend-and-Excite [150] represents an early effort in this area, introducing an attention-based Generative Semantic Nursing (GSN) mechanism. This mechanism refines cross-attention units to more effectively ensure that all subjects described in the text prompt are accurately generated. Structure Diffusion [149] employs linguistic insights to manipulate the cross-attention map, aiming for more accurate attribute binding and improved image composition. Ge *et al.* [152] propose a rich-text-to-image framework, which initially processes plain text through a diffusion model to gather attention maps, noised generation, and residual feature maps. Subsequently, rich texts are formatted as JSON to provide detailed attributes for each token span, enhancing the model's capacity to align visual content with complex textual descriptions. Additionally, SynGen [153] employs a unique methodology in text-to-image generation by first conducting a syntactic analysis of the text prompt. This analysis aims to identify entities and their modifiers within the prompt. Following this, SynGen utilizes a novel loss function designed to align the cross-attention maps with the linguistic bindings as indicated by the syntax. Furthermore, Tailored Visions [154] leverages historical user interactions with the system to rewrite user prompts to enhance the expressiveness and alignment of user prompts with their intended visual outputs. To improve textual alignment of a long paragraph (up to 512 words), Wu *et al.* [155] introduce an informative-enriched diffusion model for paragraph-to-image generation task, termed ParaDiffusion, which employ a large language model (*e.g.*, Llama V2 [221]) to encode long-form text, followed by fine-tuning with LoRA [192] to align text-image feature spaces in generation.

- **Multilingual-Guided Generation.** GlueGen [151] aligns multilingual language model (*e.g.*, XLM-Roberta [222]) with existing text-to-image models, allowing for the generation of high-quality images from captions beyond English. PEA-Diffusion [156] is a proposed simple plug-and-play language transfer method based on knowledge distillation, where a lightweight MLP-like parameter efficient adapter with only 6M parameters is trained under teacher knowledge distillation along with a small parallel data corpus.

#### 4.4 In-Context Generation

The in-context generation task involves understanding and performing specific tasks on new query images based on a pair of task-specific example images and text guidance.

Wang *et al.* [157] introduced Prompt Diffusion, a novel approach that is jointly trained over multiple tasks using in-context prompts. This method has shown impressive results in high-quality in-context generation for trained tasks and effectively generalizes to new, unseen vision tasks with relevant prompts. Building upon this, Chen *et al.* [158] further enhance Prompt Diffusion by incorporating a vision encoder-modulated text encoder. This innovation addresses several challenges, including costly pre-training, restrictive problem

formulations, limited visual comprehension, and insufficient generalizability to out-of-distribution tasks. Moreover, Nadjenkoska propose a novel framework that separates the encoding of the visual context and preserving the structure of the query images. This results in the ability to learn from the visual context and text prompts, but also from either one of them.

#### 4.5 Brain-Guided Generation

The brain-guided generation tasks focus on controlling image creation directly from brain activities, such as electroencephalogram (EEG) recordings and functional magnetic resonance imaging (fMRI), bypassing the need to translate thoughts into text. Early studies in this domain have employed Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) to reconstruct visual images from brain signals [223]–[226]. More recently, advancements have been made with the adoption of visual diffusion models, offering enhanced capabilities in accurately translating complex brain activities into coherent visual representations [159]–[165].

Chen *et al.* [159] present a Sparse Masked Brain Modeling with Doubled-Conditioned Latent Diffusion Model (MinD-Vis) for human vision decoding. They first learn an effective self-supervised representation of fMRI data using mask modeling and then augment latent diffusion model with double-conditioning. MindDiffuser [162] is also a two-stage image reconstruction model. In the first stage, the VQ-VAE latent representations and the CLIP text embeddings decoded from fMRI are put into the image-to-image process of Stable Diffusion, which yields a preliminary image that contains semantic and structural information. Then, it utilizes the low-level CLIP visual features decoded from fMRI as supervisory information, and continually adjust the two features in the first stage through backpropagation to align the structural information.

While the above methods reconstruct visual results from fMRI, some approaches choose electroencephalogram (EEG) [164], [165], which is a non-invasive and low-cost method of recording electrical activity in the brain. DreamDiffusion [164] leverages pre-trained text-to-image models and employs temporal masked signal modeling to pre-train the EEG encoder for effective and robust EEG representations. Additionally, the method further leverages a CLIP image encoder to provide extra supervision to better align EEG, text, and image embeddings with limited EEG-image pairs.

#### 4.6 Sound-Guided Generation

GlueGen [151] aligns multi-modal encoders such as Audio-CLIP with the Stable Diffusion model, enabling sound-to-image generation. Yang *et al.* [166] propose a unified framework “Align, Adapt, and Inject” (AAI) for sound-guided image generation, editing, and stylization. In particular, this method adapts input sound into a sound token, like an ordinary word, which can plug and play with existing powerful diffusion-based Text-to-Image models.

#### 4.7 Text Rendering

The task of text rendering within synthesized images is pivotal, especially given the widespread application of text in various visual forms like posters, book covers, and memes.

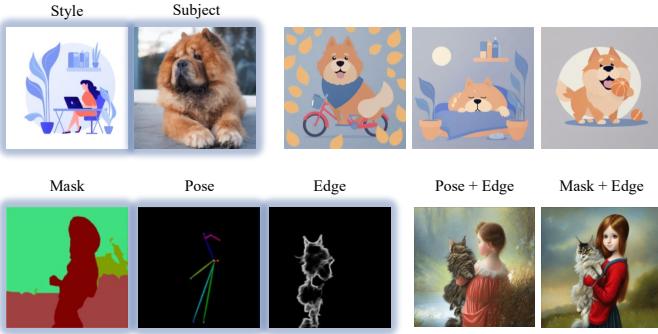


Fig. 6: **Illustration of multi-conditioned generation.** The condition is marked in blue background. Examples are sourced from [176], [182].

Drawing inspiration from the analysis in unCLIP [38], which highlights the inadequacy of raw CLIP text embeddings in accurately modeling the spelling information in prompts, subsequent efforts such as eDiff-I [26] and Imagen [24] have sought to harness the capabilities of large language models like T5 [37], trained on text-only corpora, as text encoders in image generation. Additionally, DeepFloyd IF<sup>2</sup>, following the design principles of Imagen [24], has demonstrated impressive proficiency in rendering legible text on images, showcasing a significant advancement in this challenging domain.

Meanwhile, some approaches are designed to improve text rendering capability for existing text-to-image diffusion models [167]–[169], [171]–[174]. Liu *et al.* [167] find popular text-to-image models lack character-level input features, making it much harder to predict a word’s visual makeup as a series of glyphs. GlyphControl [170] leverages additional glyph conditional information to enhance the performance of the off-the-shelf Stable-Diffusion model in generating accurate visual text. TextDiffuser *et al.* [169] first generates the layout of keywords extracted from text prompts and then generates images conditioned on the text prompt and the generated layout. The authors also contribute a large-scale text images dataset with OCR annotations, MARIO-10M, containing 10 million image-text pairs with text recognition, detection, and character-level segmentation annotations. Zhang *et al.* [174] proposed Diff-Text, a training-free scene text generation framework for any language. Diff-text leverages rendered sketch images as priors to render text by ControlNet [190] and propose a localized attention constraint to address the unreasonable position problem of scene text.

## 5 CONTROLLABLE GENERATION WITH MULTIPLE CONDITIONS

The multi-condition generation task aims to generate images under multiple conditions, such as generating a specific person in a user-defined pose or generating people in three personalized identities. In this section, we conduct a comprehensive overview of these methods from a technical perspective, categorizing them into joint training (Section 5.1),

weight fusion (Section 5.3), attention-based integration (Section 5.4), guidance fusion (Section 5.5), and continual learning (Section 5.2). Note that some of the other controllable generation methods also demonstrate multi-condition synthesis capability without dedicated designs [41], [42], [215].

### 5.1 Joint Training

Designing a multi-condition framework and jointly training them is a simple yet effective route to realize multi-condition generation. These methods generally focus on multi-condition encoders and training strategies.

Composer [175] projects all conditions (including text caption, depthmap, sketch, and *etc.*) into uniform-dimensional embeddings with the same spatial size as the noisy latent using stacked convolutional layers. It leverages a joint training strategy to generate images from a set of representations, where it uses an independent dropout probability of 0.5 for each condition, a probability of 0.1 for dropping all conditions, and a probability of 0.1 for retaining all conditions. Additionally, Cocktail [176] proposes the controllable normalization method (ControlNorm), which has an additional layer to generate two sets of learnable parameters conditioned on all modalities. These two sets of parameters are used to fuse the external conditional signals and the original signals.

From a data perspective, SVDiff [51] utilizes a cut-mix-unmix mechanism for a multi-subject generation. It augments multi-concept data by a CutMix-like data augmentation and rewrites the correspondence text prompt. It also leverages an unmix regularization on cross-attention maps, ensuring text embeddings are only effective in the correspondence areas. This attention map constraint mechanism is also applied in FastComposer [83].

### 5.2 Continual Learning

Continual learning methods are generally proposed to address knowledge “catastrophic forgetting” in training-based conditional score prediction works. Specifically, C-LoRA [177] is composed of a continually self-regularized LoRA in cross-attention layers. It utilizes the past LoRA weight deltas to regulate the new LoRA weight deltas by guiding which parameters are most available to be updated for continual concept learning. Moreover, L<sup>2</sup>DM [178] devises a task-aware memory enhancement module and an elastic-concept distillation module, which could respectively safeguard the knowledge of both prior concepts and each past personalized concept. It utilizes a rainbow-memory bank strategy to manage long-term and short-term memory and provide regularization samples to safeguard the knowledge in the personalization process. During training, the authors further propose a concept attention artist module and orthogonal attention artist module to update noise latent for better performance. STAMINA [179] introduces forgetting-regularization and sparsity-regularization in continual learning, avoiding forgetting learned concepts and ensuring no cost to storage or inference.

### 5.3 Weight Fusion

In the realm of adapting T2I diffusion models to novel conditions via fine-tuning, weight fusion presents itself

2. <https://github.com/deep-floyd/IF>

as an intuitive approach for merging multiple conditions. These methods focus on achieving a cohesive blend of weights that incorporates each condition while ensuring that the controllability of individual conditions is retained. The goal is to seamlessly integrate various conditional aspects into a unified model, thereby enhancing its versatility and applicability across diverse scenarios. This requires a delicate balance between maintaining the integrity of each condition's influence and achieving an effective overall synthesis.

Since personalized conditions usually represent UNet's weight or text embeddings, weight fusion is an intuitive and effective way to generate images under multiple personalized conditions. Specifically, Cones [180] further fine-tunes the concept neurons after personalization for better generation quality and multi-subject generating capability. Custom Diffusion [45] introduces a constrained optimization method to merge fine-tuned key and value matrices, as follows:

$$\begin{aligned} \hat{W} &= \arg \min_W \|WC_{\text{reg}} - W_0C_{\text{reg}}\|_F \\ \text{s.t. } WC^T &= V, \text{ where } C = [c_1 \dots c_N]^T \\ \text{and } V &= [W_1c_1^T \dots W_Nc_N^T]^T \end{aligned} \quad (13)$$

where  $\{W_{n,l}^k, W_{n,l}^v\}_{n=1}^N$  represent the corresponding updated key and value matrices for added  $N$  concepts and  $C_{\text{reg}}$  is a randomly sampled text features for regularization. The objective of Equation 13 is intuitively designed to ensure that the words in the target captions are consistently aligned with the values derived from the concept matrices that have undergone fine-tuning. Similarly, Mix-of-Show [181] introduces the gradient fusion, updating weight  $W$  by  $W = \arg \min_W \sum_{i=1}^n \|(W_0 + \Delta W_i)X_i - WX_i\|_F^2$  where  $X_i$  represents the input activation of the  $i$ -th concept, and  $\|\cdot\|_F$  denotes the Frobenius norm. To integrate subject-centric and style-centric conditions, ZipLoRA [182] merges LoRA-style weights by minimizing the difference between subject/style images generated by the mixed and original LoRA models and the cosine similarity between the columns of content and style LoRAs. Po *et al.* [183] present orthogonal adaption to replace LoRA in fine-tuning, encouraging the customized models to have orthogonal residual weights for efficient fusion.

#### 5.4 Attention-based Integration

Attention-based integration methods modulate attention maps to strategically position subjects within the synthesized image, allowing for precise control over where and how each condition is represented in the final composition.

For example, Cones2 [184] edits cross-attention map by  $\text{EditedCA} \leftarrow \text{Softmax}(\text{CA} \oplus \{\eta(t) \cdot M_{s_i} | i = 1, \dots, N\})$ , where  $\oplus$  denotes the operation that adds the corresponding dimension of cross-attention map  $\text{CA}$  and pre-defined layout  $M$  and  $\eta(t)$  is a concave function controlling the edit intensity at different timestep  $t$ . Similarly, Mix-of-Show [181] employs a regionally controllable sampling method, integrating global prompt and multiple regional prompts with pre-defined masks in cross-attention.

#### 5.5 Guidance Composition

Guidance composition is an integration mechanism for synthesizing images under multiple conditions, integrating

the independent denoising results of each condition. This process is mathematically represented as:

$$\hat{\epsilon}(z_t, c_1, \dots, c_N) = \sum_{i=1}^K w_i \cdot \mathcal{M}_i \cdot \epsilon(z_t, c_i) \quad (14)$$

where  $\epsilon(z_t, c_i)$  denotes the guidance of each condition, while  $w_i$  and  $\mathcal{M}_i$  are the respective weights and spatial mask used to integrate these results.

To integrate multiple concepts, Decompose and Re-align [185] obtains the corresponding  $\mathcal{M}_i$  by their cross-attention map. Similarly, Face-diffuser [89] presents a saliency-adaptive noise fusion method to combine results from a text-driven diffusion model and a proposed subject-augmented diffusion model. Besides, Cao *et al.* [119] proposes the generalized classifier-free guidance (GCFG) for concept-centric personalization and integrates concept guidance and control guidance by manually setting intensities  $w_i$ .

## 6 UNIVERSAL CONTROLLABLE TEXT-TO-IMAGE GENERATION

Beyond approaches tailored to specific types of conditions, there exist universal methods designed to accommodate arbitrary conditions in image generation. These methods are broadly categorized into two groups based on their theoretical foundations: universal conditional score prediction framework and universal condition-guided score estimation.

### 6.1 Universal Conditional Score Prediction Framework

Universal conditional score prediction framework involves creating a framework capable of encoding any given conditions and utilizing them to predict the noise at each timestep during the image synthesis process. This approach provides a universal solution that adapts flexibly to diverse conditions. By integrating the conditional information directly into the generative model, this method allows for the dynamic adaptation of the image generation process in response to a wide array of conditions, making it versatile and applicable to various image synthesis scenarios.

DiffBlender [186] is proposed to incorporate conditions from diverse types of modalities. It categorizes conditions into multiple types to employ different techniques for guiding generation. First, image-form conditions, which contain spatially rich information, are injected in ResNet Blocks [227]. Then, spatial conditions, including grounding box and keypoints, are passed through a local self-attention module to accurately locate the desired positions of synthesized results. Moreover, non-spatial conditions like color palette and style are concatenated with textual tokens through a global self-attention module and then fed into cross-attention layers. Additionally, Emu2 [187] leverages a large generative multimodal model with 37 billion parameters for task-agnostic in-context learning to construct a universal controllable T2I generation framework. After trained on a mix of high-quality datasets, it is capable of accepting a mixture of conditions like text, locations, and image as input, and generating images in context.

## 6.2 Universal Condition-Guided Score Estimation

Other approaches utilize condition-guided score estimation to incorporate various conditions into the text-to-image diffusion models. The primary challenge lies in obtaining condition-specific guidance from the latent during the denoising process.

Universal Guidance [124] observes that the reconstructed clean image proposed in the denoising diffusion implicit model (DDIM) [228] is appropriate for a generic guidance function to provide informative feedback to guide the image generation. Given any condition  $c$  and off-the-shelf predictor  $f$ , the denoising process is guided by:

$$\hat{\epsilon}_\theta(z_t, t) = \epsilon_\theta(z_t, t) + s(t) \cdot \nabla_{z_t} \mathcal{L}(c, f(\hat{z}_0)) \quad (15)$$

where  $\hat{z}_0$  is the predicted clean image following [228]:

$$\hat{z}_0 = \frac{z_t - (\sqrt{1 - \alpha_t})\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}} \quad (16)$$

UG employs various predictors, including CLIP [39] (for text or style conditions), segmentation network [229] (for segmentation map conditions), face recognition model [230], [231] (for identity conditions), and object detector [232] (for bounding box conditions), in experiments to exhibit conditional generation capabilities with various conditions.

Similar to Universal Guidance [124], FreeDom [127] leverages off-the-shelf predictors to construct time-independent energy functions to guide the generation process. It also develops the efficient time-travel strategy, taking the current intermediate result  $z_t$  back by  $j$  steps to  $z_{t+j}$  and resampling it to the  $t$ -th timestep. This mechanism solves the problem of misalignment with conditions on large data domains, *e.g.* ImageNet [233].

While above mentioned condition-guided sampling approaches leverage off-the-shelf models and one-step estimation procedure to predict condition-related conditions, Pan *et al.* [102] present Symplecit Adjoint Guidance (SAG) in two inner stages, where SAG first estimate the clean image via  $n$  function calls and then uses the symplectic adjoint method to obtain the gradients accurately.

## 7 APPLICATIONS

In this section, we focus on innovative methods that utilize novel conditions in the generation process to address specific tasks. By emphasizing these pioneering approaches, we aim to highlight how conditional generation is not only reshaping the landscape of content creation but also broadening the horizons of creativity and functionality in various fields. The subsequent discussions will provide insights into the transformative impact of these models and their potential in diverse applications.

### 7.1 Image Manipulation

Advancements in the control of pre-trained text-to-image diffusion models have allowed for more versatile image editing techniques. For instance, inspired by DreamBooth [42], SINE [234] constructs the text prompt for fine-tuning the pretrained text-to-image model by the source image as “a photo/painting of a [\*] [class]” and edits the image by a novel model-based classifier-free guidance. Moreover, the

versatility of control conditions further enhances the editing process by integrating conditions beyond mere text. For example, Choi *et al.* [235] customize the diffusion model to employ specific elements from the reference image as editing criteria, such as substituting the cat in the source image with the cat appearance in the reference image. Additionally, several approaches [236] utilize spatial manipulation by spatial control, employing sketches or layouts to intuitively adjust the arrangement of elements within an image.

### 7.2 Image Completion and Inpainting

The advancement of flexible control mechanisms has also significantly expanded the capabilities in the field of image inpainting and completion. Specifically, DreamInpainter [237] utilizes a subject-driven generation approach to personalize the filling of masked areas with the aid of reference images. Besides, Realfill [238] takes similar methods that employ reference images to facilitate realistic and coherent image completions. Moreover, by multiple condition controlling, Uni-inpaint [239] integrates a diverse set of control conditions such as text descriptions, strokes, and exemplar images to simultaneously direct the generation within the masked regions.

### 7.3 Image Composition

Image composition is a challenging task that involves multiple complex image process stages like color harmonization, geometry correction, shadow generation, and so on. While the strong prior in large-scale pre-trained diffusion model can address the problem in a unified manner. Through adding adapters to control the pre-trained text-to-image diffusion model, ObjectStitch [240] presents an object composition framework that can handle multiple aspects such as viewpoint, geometry, lighting, and shadow. Moreover, DreamCom [241] customizes the text-to-image model on several foreground object images to enhance the object details’ preservability. Besides, by inserting the task indicator vector into U-Net to control the generating process, ControlCom [242] proposes a controllable image composition method that unifies four composition-related tasks with an indicator vector.

### 7.4 Text/Image-to-3D Generation

Text/image-to-3D task aims to reconstruct 3D representations from text descriptions or images (pairs). In the early days, the task developed slowly due to their reliance on expensive 3D annotations. The strong open-world knowledge of large-scale text-to-image diffusion models brings effective solutions without heavy 3D annotation requirements. For instance, by personalizing latent diffusion model [23], Zero-1-to-3 [243] builds a viewpoint-conditioned image translation diffusion model that generates multiple views of the input object image. Then the paired images are fed into a NeRF [244] model to do reconstruction.

Recent advancements in text/image-to-3D generation represent a significant milestone with the development of Score Distillation Sampling (SDS) loss. This innovative approach, introduced by DreamFusion [245], marks a successful adaptation of large-scale 2D diffusion models for 3D

generation. Through SDS, the control method of the text-to-image model can be transferred to text-to-3D generation. Typically, DreamBooth3D [246] combines DreamBooth [42] and DreamFusion [245] that personalizes text-to-3d generative models from a few captured images of a subject. Similarly, some approaches [247], [248], [248], [249] adapt ControlNet [190] to the SDS process, enabling the control of 3D generation through spatial signals (*e.g.*, depth map, sketch).

## 8 CONCLUSION

In this comprehensive survey, we delve into the realm of conditional generation with text-to-image diffusion models, unveiling the novel conditions incorporated in the text-guided generation process. Initially, we equip readers with foundational knowledge, introducing the denoising diffusion probability models, prominent text-to-image diffusion models, and a well-structured taxonomy. Subsequently, we reveal the mechanisms of introducing novel conditions into T2I diffusion models. Then, we present a summary of previous conditional generation methods and analyze them in terms of theoretical foundations, technical advancements, and solution strategies. Furthermore, we explore the practical applications of controllable generation, underscoring its vital role and immense potential in the era of AI-generated content. This survey aims to provide a comprehensive understanding of the current landscape of controllable T2I generation, thereby contributing to the ongoing evolution and expansion of this dynamic research area.

## REFERENCES

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. [1](#)
- [2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223. [1](#)
- [3] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017. [1](#)
- [4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410. [1](#)
- [5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119. [1](#)
- [6] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021. [1](#)
- [7] P. Cao, L. Yang, D. Liu, Z. Liu, S. Li, and Q. Song, "Lsap: Rethinking inversion fidelity, perception and editability in gan latent space," *arXiv preprint arXiv:2209.12746*, 2022. [1](#)
- [8] ——, "What decreases editing capability? domain-specific hybrid refinement for improved gan inversion," *arXiv preprint arXiv:2301.12141*, 2023. [1](#)
- [9] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017. [1](#)
- [10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V* 13. Springer, 2014, pp. 740–755. [1, 9](#)
- [11] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016. [1](#)
- [12] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565. [1](#)
- [13] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3558–3568. [1](#)
- [14] K. Desai, G. Kaul, Z. Aysola, and J. Johnson, "Redcaps: Web-curated image-text data created by the people, for the people," *arXiv preprint arXiv:2111.11431*, 2021. [1](#)
- [15] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 2443–2449. [1](#)
- [16] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021. [1, 4](#)
- [17] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294, 2022. [1, 4](#)
- [18] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021. [1, 2, 3, 4](#)
- [19] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=PxTIG12RRHS> [1, 3](#)
- [20] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022. [1, 3, 4](#)
- [21] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021. [1, 2, 4](#)
- [22] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831. [1, 3, 4](#)
- [23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695. [1, 3, 4, 14](#)
- [24] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36479–36494, 2022. [1, 3, 4, 12](#)
- [25] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sndl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023. [1, 4](#)
- [26] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro *et al.*, "Ediffi: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022. [1, 5, 10, 12](#)
- [27] L. Chen, M. Zhao, Y. Liu, M. Ding, Y. Song, S. Wang, X. Wang, H. Yang, J. Liu, K. Du *et al.*, "Photoverse: Tuning-free image customization with text-to-image diffusion models," *arXiv preprint arXiv:2309.05793*, 2023. [1, 5, 7, 8](#)
- [28] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023. [2](#)
- [29] A. Ulhaq, N. Akhtar, and G. Pogrebna, "Efficient diffusion models for vision: A survey," *arXiv preprint arXiv:2210.09292*, 2022. [2](#)
- [30] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing, "Multimodal image synthesis and

- editing: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [31] F.-A. Croitoru, V. Hondu, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [32] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative ai: A survey," *arXiv preprint arXiv:2303.07909*, 2023. 2, 4
- [33] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, "A survey on video diffusion models," *arXiv preprint arXiv:2310.10647*, 2023. 2
- [34] Anonymous, "Video diffusion models - a survey," *Submitted to Transactions on Machine Learning Research*, 2023, under review. [Online]. Available: <https://openreview.net/forum?id=sgDFqNTdaN2>
- [35] C. Li, C. Zhang, A. Waghwase, L.-H. Lee, F. Rameau, Y. Yang, S.-H. Bae, and C. S. Hong, "Generative ai meets 3d: A survey on text-to-3d in aigc era," *arXiv preprint arXiv:2305.06131*, 2023. 2
- [36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 3, 4
- [37] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020. 3, 4, 12
- [38] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022. 3, 4, 9, 12
- [39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 3, 4, 8, 14
- [40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. 3, 4
- [41] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022. 5, 6, 12
- [42] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22500–22510. 5, 6, 7, 12, 14, 15
- [43] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," *arXiv preprint arXiv:2209.14491*, 2022. 5, 8
- [44] Z. Dong, P. Wei, and L. Lin, "Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning," *arXiv preprint arXiv:2211.11337*, 2022. 5, 6
- [45] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1931–1941. 5, 6, 7, 10, 13
- [46] A. Voronov, M. Khoroshikh, A. Babenko, and M. Ryabinin, "Is this loss informative? faster text-to-image customization by tracking objective dynamics," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5, 7
- [47] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Designing an encoder for fast personalization of text-to-image models," *arXiv preprint arXiv:2302.12228*, 2023. 5, 8
- [48] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, and W. Zuo, "Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation," *arXiv preprint arXiv:2302.13848*, 2023. 5, 7, 8
- [49] Y. Ma, H. Yang, W. Wang, J. Fu, and J. Liu, "Unified multimodal latent diffusion for joint subject and text conditional image generation," *arXiv preprint arXiv:2303.09319*, 2023. 5, 8
- [50] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman, "p+: Extended textual conditioning in text-to-image generation," *arXiv preprint arXiv:2303.09522*, 2023. 5, 6
- [51] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, and F. Yang, "Svdiff: Compact parameter space for diffusion fine-tuning," *arXiv preprint arXiv:2303.11305*, 2023. 5, 7, 12
- [52] C. Xiang, F. Bao, C. Li, H. Su, and J. Zhu, "A closer look at parameter-efficient tuning in diffusion models," *arXiv preprint arXiv:2303.18181*, 2023. 5, 6
- [53] W. Chen, H. Hu, Y. Li, N. Rui, X. Jia, M.-W. Chang, and W. W. Cohen, "Subject-driven text-to-image generation via apprenticeship learning," *arXiv preprint arXiv:2304.00186*, 2023. 5, 8
- [54] X. Jia, Y. Zhao, K. C. Chan, Y. Li, H. Zhang, B. Gong, T. Hou, H. Wang, and Y.-C. Su, "Taming encoder for zero fine-tuning image customization with text-to-image diffusion models," *arXiv preprint arXiv:2304.02642*, 2023. 5
- [55] J. Shi, W. Xiong, Z. Lin, and H. J. Jung, "Instantbooth: Personalized text-to-image generation without test-time finetuning," *arXiv preprint arXiv:2304.03411*, 2023. 5, 8
- [56] J. Yang, H. Wang, R. Xiao, S. Wu, G. Chen, and J. Zhao, "Controllable textual inversion for personalized text-to-image generation," *arXiv preprint arXiv:2304.05265*, 2023. 5, 7, 8
- [57] Z. Fei, M. Fan, and J. Huang, "Gradient-free textual inversion," *arXiv preprint arXiv:2304.05818*, 2023. 5, 7
- [58] Y. Tewel, R. Gal, G. Chechik, and Y. Atzmon, "Key-locked rank one editing for text-to-image personalization," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11. 5, 7
- [59] H. Chen, Y. Zhang, X. Wang, X. Duan, Y. Zhou, and W. Zhu, "Disenbooth: Disentangled parameter-efficient tuning for subject-driven text-to-image generation," *arXiv preprint arXiv:2305.03374*, 2023. 5, 6, 7
- [60] D. Li, J. Li, and S. C. Hoi, "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," *arXiv preprint arXiv:2305.14720*, 2023. 5, 8
- [61] Y. Zhang, W. Dong, F. Tang, N. Huang, H. Huang, C. Ma, T.-Y. Lee, O. Deussen, and C. Xu, "Prospect: Prompt spectrum for attribute-aware personalization of diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–14, 2023. 5, 6
- [62] O. Avrahami, K. Aberman, O. Fried, D. Cohen-Or, and D. Lischinski, "Break-a-scene: Extracting multiple concepts from a single image," *arXiv preprint arXiv:2305.16311*, 2023. 5, 6, 7
- [63] J. Xiao, M. Yin, Y. Gong, X. Zang, J. Ren, and B. Yuan, "Comcat: Towards efficient compression and customization of attention-based vision models," *arXiv preprint arXiv:2305.17235*, 2023. 5, 6
- [64] Z. Qiu, W. Liu, H. Feng, Y. Xue, Y. Feng, Z. Liu, D. Zhang, A. Weller, and B. Schölkopf, "Controlling text-to-image diffusion by orthogonal finetuning," *arXiv preprint arXiv:2306.07280*, 2023. 5, 7
- [65] Y. Li, H. Liu, Y. Wen, and Y. J. Lee, "Generate anything anywhere in any scene," *arXiv preprint arXiv:2306.17154*, 2023. 5, 6, 7
- [66] M. Arar, R. Gal, Y. Atzmon, G. Chechik, D. Cohen-Or, A. Shamir, and A. H. Bermano, "Domain-agnostic tuning-encoder for fast personalization of text-to-image models," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–10. 5, 8
- [67] J. Ma, J. Liang, C. Chen, and H. Lu, "Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning," *arXiv preprint arXiv:2307.11410*, 2023. 5, 8
- [68] S.-Y. Yeh, Y.-G. Hsieh, Z. Gao, B. B. Yang, G. Oh, and Y. Gong, "Navigating text-to-image customization: From lycoris fine-tuning to model evaluation," *arXiv preprint arXiv:2309.14859*, 2023. 5, 6
- [69] X. Pan, L. Dong, S. Huang, Z. Peng, W. Chen, and F. Wei, "Kosmos-g: Generating images in context with multimodal large language models," *arXiv preprint arXiv:2310.02992*, 2023. 5, 8
- [70] C. Jin, R. Tanno, A. Saseendran, T. Diethe, and P. Teare, "An image is worth multiple words: Learning object level concepts using multi-concept prompt learning," *arXiv preprint arXiv:2310.12274*, 2023. 5, 7
- [71] X. He, Z. Cao, N. Kolkin, L. Yu, H. Rhodin, and R. Kalarot, "A data perspective on enhanced identity preservation for diffusion personalization," *arXiv preprint arXiv:2311.04315*, 2023. 5, 7, 8
- [72] A. Roy, M. Suin, A. Shah, K. Shah, J. Liu, and R. Chellappa, "Diffnat: Improving diffusion image quality using natural image statistics," *arXiv preprint arXiv:2311.09753*, 2023. 5, 6
- [73] A. Agarwal, S. Karanam, T. Shukla, and B. V. Srinivasan, "An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis," *arXiv preprint arXiv:2311.11919*, 2023. 5, 7, 10
- [74] S. Motamed, D. P. Paudel, and L. Van Gool, "Lego: Learning to disentangle and invert concepts beyond object appearance in text-to-image diffusion models," *arXiv preprint arXiv:2311.13833*, 2023. 5, 6

- [75] R. Zhao, M. Zhu, S. Dong, N. Wang, and X. Gao, "Catversion: Concatenating embeddings for diffusion-based text-to-image personalization," *arXiv preprint arXiv:2311.14631*, 2023. [5](#), [6](#), [7](#)
- [76] M. Safaei, A. Mikaeili, O. Patashnik, D. Cohen-Or, and A. Mahdavi-Amiri, "Clic: Concept learning in context," *arXiv preprint arXiv:2311.17083*, 2023. [5](#), [7](#)
- [77] H. Zhao, T. Lu, J. Gu, X. Zhang, Z. Wu, H. Xu, and Y.-G. Jiang, "Videoassembler: Identity-consistent video generation with reference entities using diffusion model," *arXiv preprint arXiv:2311.17338*, 2023. [5](#), [6](#), [8](#)
- [78] Z. Wang, W. Wei, Y. Zhao, Z. Xiao, M. Hasegawa-Johnson, H. Shi, and T. Hou, "Hifi tuner: High-fidelity subject-driven fine-tuning for diffusion models," *arXiv preprint arXiv:2312.00079*, 2023. [5](#), [7](#), [8](#)
- [79] Y. Jiang, T. Wu, S. Yang, C. Si, D. Lin, Y. Qiao, C. C. Loy, and Z. Liu, "Videobooth: Diffusion-based video generation with image prompts," *arXiv preprint arXiv:2312.00777*, 2023. [5](#), [8](#)
- [80] Y. Zhou, R. Zhang, J. Gu, and T. Sun, "Customization assistant for text-to-image generation," *arXiv preprint arXiv:2312.03045*, 2023. [5](#), [8](#)
- [81] Y. Cai, Y. Wei, Z. Ji, J. Bai, H. Han, and W. Zuo, "Decoupled textual embeddings for customized image generation," *arXiv preprint arXiv:2312.11826*, 2023. [5](#), [6](#), [7](#)
- [82] M. Hua, J. Liu, F. Ding, W. Liu, J. Wu, and Q. He, "Dreamtuner: Single image is enough for subject-driven generation," 2023. [5](#), [7](#), [8](#), [10](#)
- [83] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han, "Fastcomposer: Tuning-free multi-subject image generation with localized attention," *arXiv preprint arXiv:2305.10431*, 2023. [5](#), [8](#), [12](#)
- [84] N. Giambi and G. Lisanti, "Conditioning diffusion models via attributes and semantic masks for face generation," *arXiv preprint arXiv:2306.00914*, 2023. [5](#), [10](#)
- [85] D. Valembois, D. Lumen, Y. Matias, and Y. Leviathan, "Face0: Instantaneously conditioning a text-to-image model on a face," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–10. [5](#), [8](#)
- [86] Z. Chen, S. Fang, W. Liu, Q. He, M. Huang, Y. Zhang, and Z. Mao, "Dreamidentity: Improved editability for efficient face-identity preserved image generation," *arXiv preprint arXiv:2307.00300*, 2023. [5](#), [8](#)
- [87] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, and K. Aberman, "Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models," *arXiv preprint arXiv:2307.06949*, 2023. [5](#), [8](#), [10](#)
- [88] J. Hyung, J. Shin, and J. Choo, "Magicapture: High-resolution multi-concept portrait customization," *arXiv preprint arXiv:2309.06895*, 2023. [5](#), [8](#)
- [89] Y. Wang, W. Zhang, J. Zheng, and C. Jin, "High-fidelity person-centric subject-to-image synthesis," *arXiv preprint arXiv:2311.10329*, 2023. [5](#), [13](#)
- [90] X. Li, X. Hou, and C. C. Loy, "When stylegan meets stable diffusion: a  $\mathbb{W}$ -adapter for personalized image generation," *arXiv preprint arXiv:2311.17461*, 2023. [5](#), [8](#)
- [91] H. Tang, X. Zhou, J. Deng, Z. Pan, H. Tian, and P. Chaudhari, "Retrieving conditions from reference images for diffusion models," *arXiv preprint arXiv:2312.02521*, 2023. [5](#), [8](#)
- [92] Y. Yan, C. Zhang, R. Wang, Y. Zhou, G. Zhang, P. Cheng, G. Yu, and B. Fu, "Facestudio: Put your face everywhere in seconds," *arXiv preprint arXiv:2312.02663*, 2023. [5](#)
- [93] S. Y. Cheong, A. Mustafa, and A. Gilbert, "Visconet: Bridging and harmonizing visual and textual conditioning for controlnet," *arXiv preprint arXiv:2312.03154*, 2023. [5](#), [9](#)
- [94] D.-Y. Chen, S. Koley, A. Sain, P. N. Chowdhury, T. Xiang, A. K. Bhunia, and Y.-Z. Song, "Democaricature: Democratising caricature generation with a rough sketch," *arXiv preprint arXiv:2312.04364*, 2023. [5](#)
- [95] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, "Photomaker: Customizing realistic human photos via stacked id embedding," *arXiv preprint arXiv:2312.04461*, 2023. [5](#), [8](#)
- [96] P. Achlioptas, A. Benetatos, I. Fostirooulos, and D. Skourtis, "Stellar: Systematic evaluation of human-centric personalized text-to-image methods," *arXiv preprint arXiv:2312.06116*, 2023. [5](#), [8](#)
- [97] X. Peng, J. Zhu, B. Jiang, Y. Tai, D. Luo, J. Zhang, W. Lin, T. Jin, C. Wang, and R. Ji, "Portraitbooth: A versatile portrait model for fast identity-preserved personalization," *arXiv preprint arXiv:2312.06354*, 2023. [5](#), [8](#)
- [98] K. Sohn, N. Ruiz, K. Lee, D. C. Chin, I. Blok, H. Chang, J. Barber, L. Jiang, G. Entis, Y. Li *et al.*, "Styledrop: Text-to-image generation in any style," *arXiv preprint arXiv:2306.00983*, 2023. [5](#), [8](#)
- [99] G. Liu, M. Xia, Y. Zhang, H. Chen, J. Xing, X. Wang, Y. Yang, and Y. Shan, "Stylecrafter: Enhancing stylized text-to-video generation with style adapter," *arXiv preprint arXiv:2312.00330*, 2023. [5](#), [9](#)
- [100] D.-Y. Chen, H. Tennent, and C.-W. Hsu, "Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation," *arXiv preprint arXiv:2312.02109*, 2023. [5](#), [7](#), [9](#)
- [101] A. Hertz, A. Voynov, S. Fruchter, and D. Cohen-Or, "Style aligned image generation via shared attention," *arXiv preprint arXiv:2312.02133*, 2023. [4](#), [5](#), [9](#)
- [102] J. Pan, H. Yan, J. H. Liew, J. Feng, and V. Y. Tan, "Towards accurate guided diffusion sampling through symplectic adjoint method," *arXiv preprint arXiv:2312.12030*, 2023. [5](#), [9](#), [14](#)
- [103] Z. Huang, T. Wu, Y. Jiang, K. C. Chan, and Z. Liu, "Reversion: Diffusion-based relation inversion from images," *arXiv preprint arXiv:2303.13495*, 2023. [5](#), [9](#)
- [104] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023. [5](#)
- [105] R. Zhao, Y. Gu, J. Z. Wu, D. J. Zhang, J. Liu, W. Wu, J. Keppo, and M. Z. Shou, "Motiondirector: Motion customization of text-to-video diffusion models," *arXiv preprint arXiv:2310.08465*, 2023. [5](#)
- [106] R. Wu, L. Chen, T. Yang, C. Guo, C. Li, and X. Zhang, "Lamp: Learn a motion pattern for few-shot-based video generation," *arXiv preprint arXiv:2310.10769*, 2023. [5](#)
- [107] Y. Song, W. Shin, J. Lee, J. Kim, and N. Kwak, "Save: Protagonist diversification with structure agnostic video editing," *arXiv preprint arXiv:2312.02503*, 2023. [5](#)
- [108] J. Materzynska, J. Sivic, E. Shechtman, A. Torralba, R. Zhang, and B. Russell, "Customizing motion in text-to-video diffusion models," *arXiv preprint arXiv:2312.04966*, 2023. [5](#)
- [109] M. Feng, J. Liu, K. Yu, Y. Yao, Z. Hui, X. Guo, X. Lin, H. Xue, C. Shi, X. Li *et al.*, "Dreamoving: A human dance video generation framework based on diffusion models," *arXiv preprint arXiv:2312.05107*, 2023. [5](#)
- [110] Y. Zhang, F. Tang, N. Huang, H. Huang, C. Ma, W. Dong, and C. Xu, "Motioncrafter: One-shot motion customization of diffusion models," *arXiv preprint arXiv:2312.05288*, 2023. [5](#)
- [111] J. Tian Hoe, X. Jiang, C. S. Chan, Y.-P. Tan, and W. Hu, "Interact-diffusion: Interaction control in text-to-image diffusion models," *arXiv e-prints*, pp. arXiv-2312, 2023. [5](#), [9](#), [10](#)
- [112] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022. [5](#), [7](#)
- [113] X. Xu, Z. Wang, G. Zhang, K. Wang, and H. Shi, "Versatile diffusion: Text, images and variations all in one diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7754–7765. [5](#), [9](#)
- [114] X. Xu, J. Guo, Z. Wang, G. Huang, I. Essa, and H. Shi, "Prompt-free diffusion: Taking "text" out of text-to-image diffusion models," *arXiv preprint arXiv:2305.16223*, 2023. [5](#), [9](#)
- [115] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, "Uni-controlnet: All-in-one control to text-to-image diffusion models," *arXiv preprint arXiv:2305.16322*, 2023. [5](#), [10](#)
- [116] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023. [5](#), [9](#)
- [117] I. Najdenkoska, A. Sinha, A. Dubey, D. Mahajan, V. Ramanathan, and F. Radenovic, "Context diffusion: In-context aware image generation," *arXiv preprint arXiv:2312.03584*, 2023. [5](#)
- [118] S. Mo, F. Mu, K. H. Lin, Y. Liu, B. Guan, Y. Li, and B. Zhou, "Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition," *arXiv preprint arXiv:2312.07536*, 2023. [5](#), [10](#)
- [119] P. Cao, L. Yang, F. Zhou, T. Huang, and Q. Song, "Concept-centric personalization with large-scale diffusion priors," *arXiv preprint arXiv:2312.08195*, 2023. [5](#), [7](#), [9](#), [13](#)
- [120] B. Nlong Zhao, Y. Xiao, J. Xu, X. Jiang, Y. Yang, D. Li, L. Itti, V. Vineet, and Y. Ge, "Dreamdistribution: Prompt distribution learning for text-to-image diffusion models," *arXiv e-prints*, pp. arXiv-2312, 2023. [5](#), [9](#)
- [121] A. Voynov, K. Aberman, and D. Cohen-Or, "Sketch-guided text-to-image diffusion models," in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11. [5](#), [10](#)
- [122] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin, "Spatext: Spatio-textual

- representation for controllable image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 370–18 380. 5, 10
- [123] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 511–22 521. 5, 10
- [124] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein, "Universal guidance for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 843–852. 5, 10, 14
- [125] J. Cheng, X. Liang, X. Shi, T. He, T. Xiao, and M. Li, "Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation," *arXiv preprint arXiv:2302.08908*, 2023. 5, 10
- [126] C. Ham, J. Hays, J. Lu, K. K. Singh, Z. Zhang, and T. Hinz, "Modulating pretrained diffusion models for multimodal image synthesis," *arXiv preprint arXiv:2302.12764*, 2023. 5, 10
- [127] J. Yu, Y. Wang, C. Zhao, B. Ghanem, and J. Zhang, "Freedom: Training-free energy-guided conditional diffusion model," *arXiv preprint arXiv:2303.09833*, 2023. 5, 10, 14
- [128] H. Xue, Z. Huang, Q. Sun, L. Song, and W. Zhang, "Freestyle layout-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 256–14 266. 5, 10
- [129] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "Layoutdiffusion: Controllable diffusion model for layout-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 490–22 499. 5, 10
- [130] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q. Xu, "Humansd: A native skeleton-guided diffusion model for human image generation," *arXiv preprint arXiv:2304.04269*, 2023. 5
- [131] C. Liu and D. Liu, "Late-constraint diffusion guidance for controllable image synthesis," *arXiv preprint arXiv:2305.11520*, 2023. 5, 10
- [132] K. Chen, E. Xie, Z. Chen, L. Hong, Z. Li, and D.-Y. Yeung, "Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt," *arXiv preprint arXiv:2306.04607*, 2023. 5, 10
- [133] Q. Phung, S. Ge, and J.-B. Huang, "Grounded text-to-image synthesis with attention refocusing," *arXiv preprint arXiv:2306.05427*, 2023. 5, 10
- [134] G. Couairon, M. Careil, M. Cord, S. Lathuiliere, and J. Verbeek, "Zero-shot spatial layout conditioning for text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2174–2183. 5, 10
- [135] Y. He, R. Salakhutdinov, and J. Z. Kolter, "Localized text-to-image generation for free via cross attention control," *arXiv preprint arXiv:2306.14636*, 2023. 4, 5, 10
- [136] C. Jia, M. Luo, Z. Dang, G. Dai, X. Chang, M. Wang, and J. Wang, "Ssmg: Spatial-semantic map guided diffusion model for free-form layout-to-image generation," *arXiv preprint arXiv:2308.10156*, 2023. 5, 10
- [137] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu, "Dense text-to-image generation with attention modulation," *ArXiv*, vol. abs/2308.12964, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:261101003> 5, 10
- [138] J. Zhang, S. Li, Y. Lu, T. Fang, D. McKinnon, Y. Tsin, L. Quan, and Y. Yao, "Jointnet: Extending text-to-image diffusion for dense distribution modeling," *arXiv preprint arXiv:2310.06347*, 2023. 5, 10
- [139] X. Liu, J. Ren, A. Siarohin, I. Skorokhodov, Y. Li, D. Lin, X. Liu, Z. Liu, and S. Tulyakov, "Hyperhuman: Hyper-realistic human generation with latent structural diffusion," *arXiv preprint arXiv:2310.08579*, 2023. 5, 10
- [140] C. F. Dormann, J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolíger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. Daniel Kissling *et al.*, "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review," *Ecography*, vol. 30, no. 5, pp. 609–628, 2007. 5
- [141] Y. Wang, W. Zhang, J. Zheng, and C. Jin, "Enhancing object coherence in layout-to-image synthesis," *arXiv preprint arXiv:2311.10522*, 2023. 5, 10
- [142] P. Zhao, H. Li, R. Jin, and S. K. Zhou, "Loco: Locally constrained training-free layout-to-image synthesis," *arXiv preprint arXiv:2311.12342*, 2023. 4, 5, 10
- [143] A. Voynov, A. Hertz, M. Arar, S. Fruchter, and D. Cohen-Or, "Anylens: A generative diffusion model with any rendering lens," *arXiv preprint arXiv:2311.17609*, 2023. 5, 10
- [144] Z. Qi, G. Huang, Z. Huang, Q. Guo, J. Chen, J. Han, J. Wang, G. Zhang, L. Liu, E. Ding *et al.*, "Layered rendering diffusion model for zero-shot guided image synthesis," *arXiv preprint arXiv:2311.18435*, 2023. 5, 10
- [145] S. F. Bhat, N. J. Mitra, and P. Wonka, "Loosecontrol: Lifting controlnet for generalized depth conditioning," *arXiv preprint arXiv:2312.03079*, 2023. 5, 10
- [146] Y. Zhao, L. Peng, Y. Yang, Z. Luo, H. Li, Y. Chen, W. Zhao, W. Liu, B. Wu *et al.*, "Local conditional controlling for text-to-image diffusion models," *arXiv preprint arXiv:2312.08768*, 2023. 5, 10
- [147] Z. Jiang, C. Mao, Y. Pan, Z. Han, and J. Zhang, "Scedit: Efficient and controllable image diffusion generation via skip connection editing," *arXiv preprint arXiv:2312.11392*, 2023. 5, 10
- [148] J. Ren, C. Xu, H. Chen, X. Qin, C. Li, and L. Zhu, "Towards flexible, scalable, and adaptive multi-modal conditioned face synthesis," *arXiv preprint arXiv:2312.16274*, 2023. 5, 10
- [149] W. Feng, X. He, T.-J. Fu, V. Jampani, A. R. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, "Training-free structured diffusion guidance for compositional text-to-image synthesis," *ArXiv*, 2022. 5, 11
- [150] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–10, 2023. 5, 11
- [151] C. Qin, N. Yu, C. Xing, S. Zhang, Z. Chen, S. Ermon, Y. Fu, C. Xiong, and R. Xu, "Gluegen: Plug and play multi-modal encoders for x-to-image generation," *arXiv preprint arXiv:2303.10056*, 2023. 5, 7, 11
- [152] S. Ge, T. Park, J.-Y. Zhu, and J.-B. Huang, "Expressive text-to-image generation with rich text," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7545–7556. 5, 11
- [153] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik, "Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment," *arXiv preprint arXiv:2306.08877*, 2023. 5, 11
- [154] Z. Chen, L. Zhang, F. Weng, L. Pan, and Z. Lan, "Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting," *arXiv preprint arXiv:2310.08129*, 2023. 5, 11
- [155] W. Wu, Z. Li, Y. He, M. Z. Shou, C. Shen, L. Cheng, Y. Li, T. Gao, D. Zhang, and Z. Wang, "Paragraph-to-image generation with information-enriched diffusion model," *arXiv preprint arXiv:2311.14284*, 2023. 5, 7, 11
- [156] J. Ma, C. Chen, Q. Xie, and H. Lu, "Pea-diffusion: Parameter-efficient adapter with knowledge distillation in non-english text-to-image generation," *arXiv preprint arXiv:2311.17086*, 2023. 5, 11
- [157] Z. Wang, Y. Jiang, Y. Lu, Y. Shen, P. He, W. Chen, Z. Wang, and M. Zhou, "In-context learning unlocked for diffusion models," *arXiv preprint arXiv:2305.01115*, 2023. 5, 7, 11
- [158] T. Chen, Y. Liu, Z. Wang, J. Yuan, Q. You, H. Yang, and M. Zhou, "Improving in-context learning in diffusion models with visual context-modulated prompts," *arXiv preprint arXiv:2312.01408*, 2023. 5, 11
- [159] Z. Chen, J. Qing, T. Xiang, W. L. Yue, and J. H. Zhou, "Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 710–22 720. 5, 11
- [160] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 453–14 463. 5, 11
- [161] F. Ozcelik and R. VanRullen, "Natural scene reconstruction from fmri signals using generative latent diffusion," *Scientific Reports*, vol. 13, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:260439960> 5, 11
- [162] Y. Lu, C. Du, Q. Zhou, D. Wang, and H. He, "Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5899–5908. 5, 7, 11
- [163] P. Ni and Y. Zhang, "Natural image reconstruction from fmri based on self-supervised representation learning and latent diffusion

- model," in *Proceedings of the 15th International Conference on Digital Image Processing*, 2023, pp. 1–9. 5, 11
- [164] Y. Bai, X. Wang, Y.-p. Cao, Y. Ge, C. Yuan, and Y. Shan, "Dream-diffusion: Generating high-quality images from brain eeg signals," *arXiv preprint arXiv:2306.16934*, 2023. 5, 11
- [165] H. Fu, Z. Shen, J. J. Chin, and H. Wang, "Brainvis: Exploring the bridge between brain and visual signals via image reconstruction," *arXiv preprint arXiv:2312.14871*, 2023. 5, 11
- [166] Y. Yang, K. Zhang, Y. Ge, W. Shao, Z. Xue, Y. Qiao, and P. Luo, "Align, adapt and inject: Sound-guided unified image generation," *arXiv preprint arXiv:2306.11504*, 2023. 5, 11
- [167] R. Liu, D. Garrette, C. Saharia, W. Chan, A. Roberts, S. Narang, I. Blok, R. Mical, M. Norouzi, and N. Constant, "Character-aware models improve visual text rendering," *arXiv preprint arXiv:2212.10562*, 2022. 5, 12
- [168] J. Ma, M. Zhao, C. Chen, R. Wang, D. Niu, H. Lu, and X. Lin, "Glyphdraw: Learning to draw chinese characters in image synthesis models coherently," *arXiv preprint arXiv:2303.17870*, 2023. 5, 12
- [169] J. Chen, Y. Huang, T. Lv, L. Cui, Q. Chen, and F. Wei, "Textdiffuser: Diffusion models as text painters," *arXiv preprint arXiv:2305.10855*, 2023. 5, 12
- [170] Y. Yang, D. Gui, Y. Yuan, H. Ding, H. Hu, and K. Chen, "Glyphcontrol: Glyph conditional control for visual text generation," *arXiv preprint arXiv:2305.18259*, 2023. 5, 12
- [171] Y. Tuo, W. Xiang, J.-Y. He, Y. Geng, and X. Xie, "Anytext: Multilingual visual text generation and editing," *arXiv preprint arXiv:2311.03054*, 2023. 5, 12
- [172] J. Chen, Y. Huang, T. Lv, L. Cui, Q. Chen, and F. Wei, "Textdiffuser-2: Unleashing the power of language models for text rendering," *arXiv preprint arXiv:2311.16465*, 2023. 5, 12
- [173] Y. Zhao and Z. Lian, "Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models," *arXiv preprint arXiv:2312.04884*, 2023. 5, 12
- [174] L. Zhang, X. Chen, Y. Wang, Y. Lu, and Y. Qiao, "Brush your text: Synthesize any scene text on images via diffusion model," *arXiv preprint arXiv:2312.12232*, 2023. 5, 7, 12
- [175] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," *arXiv preprint arXiv:2302.09778*, 2023. 5, 12
- [176] M. Hu, J. Zheng, D. Liu, C. Zheng, C. Wang, D. Tao, and T.-J. Cham, "Cocktail: Mixing multi-modality controls for text-conditional image generation," *arXiv preprint arXiv:2306.00964*, 2023. 5, 10, 12
- [177] J. S. Smith, Y.-C. Hsu, L. Zhang, T. Hua, Z. Kira, Y. Shen, and H. Jin, "Continual diffusion: Continual customization of text-to-image diffusion with c-lora," *arXiv preprint arXiv:2304.06027*, 2023. 5, 6, 12
- [178] G. Sun, W. Liang, J. Dong, J. Li, Z. Ding, and Y. Cong, "Create your world: Lifelong text-to-image diffusion," *arXiv preprint arXiv:2309.04430*, 2023. 5, 12
- [179] J. S. Smith, Y.-C. Hsu, Z. Kira, Y. Shen, and H. Jin, "Continual diffusion with stamina: Stack-and-mask incremental adapters," *arXiv preprint arXiv:2311.18763*, 2023. 5, 12
- [180] Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao, "Cones: Concept neurons in diffusion models for customized generation," *arXiv preprint arXiv:2303.05125*, 2023. 5, 13
- [181] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu et al., "Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models," *arXiv preprint arXiv:2305.18292*, 2023. 5, 6, 13
- [182] V. Shah, N. Ruiz, F. Cole, E. Lu, S. Lazebnik, Y. Li, and V. Jampani, "Ziplora: Any subject in any style by effectively merging loras," *arXiv preprint arXiv:2311.13600*, 2023. 5, 12, 13
- [183] R. Po, G. Yang, K. Aberman, and G. Wetzstein, "Orthogonal adaptation for modular customization of diffusion models," *arXiv preprint arXiv:2312.02432*, 2023. 5, 13
- [184] Z. Liu, Y. Zhang, Y. Shen, K. Zheng, K. Zhu, R. Feng, Y. Liu, D. Zhao, J. Zhou, and Y. Cao, "Cones 2: Customizable image synthesis with multiple subjects," *arXiv preprint arXiv:2305.19327*, 2023. 5, 13
- [185] L. Wang, G. Shen, Y. Li, and Y.-c. Chen, "Decompose and realign: Tackling condition misalignment in text-to-image diffusion models," *arXiv preprint arXiv:2306.14408*, 2023. 5, 13
- [186] S. Kim, J. Lee, K. Hong, D. Kim, and N. Ahn, "Diffblender: Scalable and composable multimodal text-to-image diffusion models," *arXiv preprint arXiv:2305.15194*, 2023. 5, 10, 13
- [187] Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Z. Luo, Y. Wang, Y. Rao, J. Liu, T. Huang et al., "Generative multimodal models are in-context learners," *arXiv preprint arXiv:2312.13286*, 2023. 5, 13
- [188] C. Luo, "Understanding diffusion models: A unified perspective," *arXiv preprint arXiv:2208.11970*, 2022. 3
- [189] S. Huang, B. Gong, Y. Feng, X. Chen, Y. Fu, Y. Liu, and D. Wang, "Learning disentangled identifiers for action-customized text-to-image generation," *arXiv preprint arXiv:2311.15841*, 2023. 7, 9
- [190] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847. 7, 10, 12, 15
- [191] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Larosière, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799. 6, 9
- [192] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021. 6, 10, 11
- [193] M. Valipour, M. Rezagholizadeh, I. Kobyzhev, and A. Ghodsi, "Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation," *arXiv preprint arXiv:2210.07558*, 2022. 6
- [194] A. Chavan, Z. Liu, D. Gupta, E. Xing, and Z. Shen, "One-for-all: Generalized lora for parameter-efficient fine-tuning," *arXiv preprint arXiv:2306.07967*, 2023. 6
- [195] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:256390509>. 8, 9
- [196] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 1. 8
- [197] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74. 8
- [198] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017. 8, 9
- [199] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020. 8
- [200] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 8
- [201] K. Sohn, H. Chang, J. Lezama, L. Polania, H. Zhang, Y. Hao, I. Essa, and L. Jiang, "Visual prompt tuning for generative transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19840–19851. 9
- [202] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510. 9
- [203] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636. 9
- [204] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660. 9
- [205] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693. 9
- [206] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3d human pose estimation: A review," *Computer Vision and Image Understanding*, vol. 210, p. 103225, 2021. 9
- [207] F. Zhou, J. Yin, and P. Li, "Lifting by image-leveraging image cues for accurate 3d human pose estimation," *arXiv preprint arXiv:2312.15636*, 2023. 9
- [208] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-cnn for instance-level human analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 364–373. 9

- [209] L. Yang, Q. Song, Z. Wang, M. Hu, C. Liu, X. Xin, W. Jia, and S. Xu, "Renovating parsing r-cnn for accurate multiple human parsing," in *European Conference on Computer Vision*. Springer, 2020, pp. 421–437. [9](#)
- [210] L. Yang, Q. Song, X. Xin, and Z. Liu, "Quality-aware network for face parsing," *arXiv preprint arXiv:2106.07368*, 2021. [9](#)
- [211] L. Yang, Z. Liu, T. Zhou, and Q. Song, "Part decomposition and refinement network for human parsing," *IEEE/CAC Journal of Automatica Sinica*, vol. 9, no. 6, pp. 1111–1114, 2022. [9](#)
- [212] L. Yang, Q. Song, Z. Wang, Z. Liu, S. Xu, and Z. Li, "Quality-aware network for human parsing," *IEEE Transactions on Multimedia*, 2022. [9](#)
- [213] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641. [9](#)
- [214] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. [9](#)
- [215] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese *et al.*, "Unicontrol: A unified diffusion model for controllable visual generation in the wild," *arXiv preprint arXiv:2305.11147*, 2023. [10](#) [12](#)
- [216] J. Yang, J. Zhao, P. Wang, Z. Wang, and Y. Liang, "Meta controlnet: Enhancing task adaptation via meta learning," *arXiv preprint arXiv:2312.01255*, 2023. [10](#)
- [217] D. Zavadski, J.-F. Feiden, and C. Rother, "Controlnet-xs: Designing an efficient and effective architecture for controlling text-to-image diffusion models," *arXiv preprint arXiv:2312.06573*, 2023. [10](#)
- [218] J. Xiao, K. Zhu, H. Zhang, Z. Liu, Y. Shen, Y. Liu, X. Fu, and Z.-J. Zha, "Ccm: Adding conditional controls to text-to-image consistency models," *arXiv preprint arXiv:2312.06971*, 2023. [10](#)
- [219] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023. [10](#)
- [220] J. Xiao, L. Li, H. Lv, S. Wang, and Q. Huang, "R&b: Region and boundary aware zero-shot grounded text-to-image generation," *arXiv preprint arXiv:2310.08872*, 2023. [10](#)
- [221] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almohairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023. [11](#)
- [222] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019. [11](#)
- [223] I. Kavasidis, S. Palazzo, C. Spampinato, D. Giordano, and M. Shah, "Brain2image: Converting brain signals into images," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1809–1817. [11](#)
- [224] P. Tirupattur, Y. S. Rawat, C. Spampinato, and M. Shah, "Thoughtviz: Visualizing human thoughts using generative adversarial network," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 950–958. [11](#)
- [225] K. K. Bhargav, S. Ambika, S. Deepak, and S. Sudha, "Imagenation - a dcgan based method for image reconstruction from fmri," *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 112–119, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:229373783> [11](#)
- [226] S. Lin, T. Sprague, and A. K. Singh, "Mind reader: Reconstructing complex images from brain activities," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 624–29 636, 2022. [11](#)
- [227] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. [13](#)
- [228] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020. [14](#)
- [229] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324. [14](#)
- [230] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016. [14](#)
- [231] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823. [14](#)
- [232] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015. [14](#)
- [233] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255. [14](#)
- [234] Z. Zhang, L. Han, A. Ghosh, D. N. Metaxas, and J. Ren, "Sine: Single image editing with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 6027–6037. [14](#)
- [235] J. Choi, Y. Choi, Y. Kim, J. Kim, and S. Yoon, "Custom-edit: Text-guided image editing with customized diffusion models," *arXiv preprint arXiv:2305.15779*, 2023. [14](#)
- [236] Z. Zhang, Z. Huang, and J. Liao, "Continuous layout editing of single images with diffusion models," in *Computer Graphics Forum*, vol. 42, no. 7. Wiley Online Library, 2023, p. e14966. [14](#)
- [237] S. Xie, Y. Zhao, Z. Xiao, K. C. Chan, Y. Li, Y. Xu, K. Zhang, and T. Hou, "Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models," *arXiv preprint arXiv:2312.03771*, 2023. [14](#)
- [238] L. Tang, N. Ruiz, Q. Chu, Y. Li, A. Holynski, D. E. Jacobs, B. Hariharan, Y. Pritch, N. Wadhwa, K. Aberman *et al.*, "Realfill: Reference-driven generation for authentic image completion," *arXiv preprint arXiv:2309.16668*, 2023. [14](#)
- [239] S. Yang, X. Chen, and J. Liao, "Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3190–3199. [14](#)
- [240] Y. Song, Z. Zhang, Z. Lin, S. Cohen, B. Price, J. Zhang, S. Y. Kim, and D. Aliaga, "Objectstitch: Object compositing with diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 310–18 319. [14](#)
- [241] L. Lu, B. Zhang, and L. Niu, "Dreamcom: Finetuning text-guided inpainting model for image composition," *arXiv preprint arXiv:2309.15508*, 2023. [14](#)
- [242] B. Zhang, Y. Duan, J. Lan, Y. Hong, H. Zhu, W. Wang, and L. Niu, "Controlcom: Controllable image composition using diffusion model," *arXiv preprint arXiv:2308.10040*, 2023. [14](#)
- [243] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9298–9309. [14](#)
- [244] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021. [14](#)
- [245] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *The Eleventh International Conference on Learning Representations*, 2022. [14](#) [15](#)
- [246] A. Raj, S. Kaza, B. Poole, M. Niemeyer, N. Ruiz, B. Mildenhall, S. Zada, K. Aberman, M. Rubinstein, J. Barron *et al.*, "Dreambooth3d: Subject-driven text-to-3d generation," *arXiv preprint arXiv:2303.13508*, 2023. [15](#)
- [247] Y. Chen, Y. Pan, Y. Li, T. Yao, and T. Mei, "Control3d: Towards controllable text-to-3d generation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1148–1156. [15](#)
- [248] T. Huang, Y. Zeng, Z. Zhang, W. Xu, H. Xu, S. Xu, R. W. Lau, and W. Zuo, "Dreamcontrol: Control-based text-to-3d generation with 3d self-prior," *arXiv preprint arXiv:2312.06439*, 2023. [15](#)
- [249] C. Yu, Q. Zhou, J. Li, Z. Zhang, Z. Wang, and F. Wang, "Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6841–6850. [15](#)