SmartControl: Enhancing ControlNet for Handling Rough Visual Conditions

Xiaoyu Liu[1], Yuxiang Wei[1], Ming Liu[1(✉)], Xianhui Lin[2],
Peiran Ren[2], Xuansong Xie[2], and Wangmeng Zuo[1,3]

[1] Harbin Institute of Technology, Harbin, China
[2] Institute for Intelligent Computing
[3] Pazhou Lab Huangpu, Guangzhou, China
{liuxiaoyu1104,yuxiang.wei.cs}@gmail.com, csmliu@outlook.com,
{xhlin129,renpeiran,xiexuansong}@gmail.com, wmzuo@hit.edu.cn

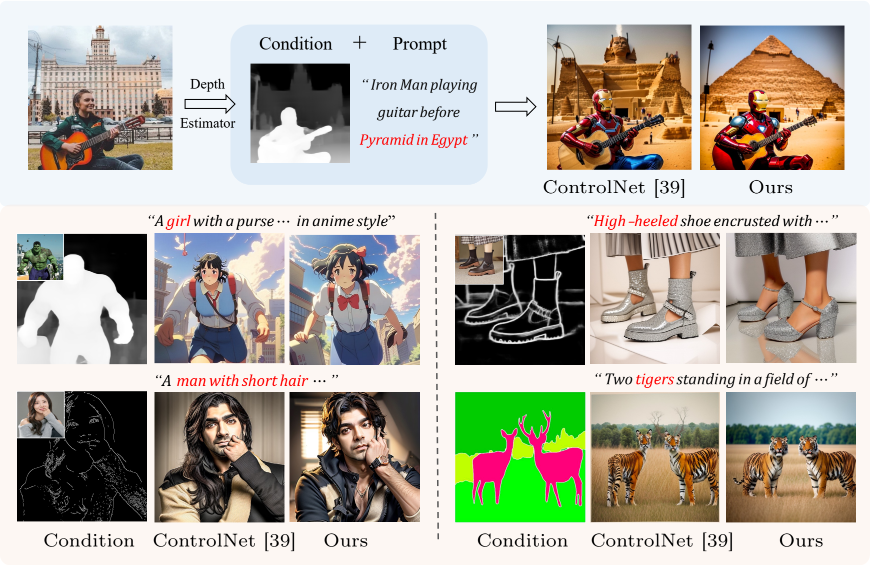
Depth
Estimator
Condition + Prompt
"Iron Man playing guitar before Pyramid in Egypt"
ControlNet [39]
Ours
"A girl with a purse ··· in anime style"
"High -heeled shoe encrusted with ···"
"A man with short hair ···"
"Two tigers standing in a field of ···"
Condition
ControlNet [39]
Ours
Condition
ControlNet [39]
Ours

Figure 1: Our proposed SmartControl can perform controllable image generation under rough visual conditions extracted from other images. In contrast, ControlNet [39] adheres strictly to control conditions, which may go against human intentions.

Abstract Recent text-to-image generation methods such as ControlNet have achieved remarkable success in controlling image layouts, where the generated images by the default model are constrained to strictly follow the visual conditions (e.g., depth maps). However, in practice, the conditions usually provide only a rough layout, and we argue that the text prompts can more faithfully reflect user intentions. For handling the disagreements between the text prompts and rough visual conditions, we propose a novel text-to-image generation method dubbed SmartControl, which is designed to align well with the text prompts while adaptively keeping useful information from the visual conditions. The key idea of our SmartControl is to relax the constraints on areas that conflict with the text prompts in visual conditions, and two main procedures are required to achieve such a flexible generation. In specific, we extract information from the generative priors of the backbone model (e.g., ControlNet), which effectively represents consistency between the text prompt and visual conditions. Then, a Control Scale Predictor is designed to identify the conflict regions and predict the local control scales. For training the proposed method, a dataset with text prompts and rough visual conditions is constructed. It is worth noting that, even with a limited number (e.g., 1,000∼2,000) of training samples, our SmartControl can generalize well to unseen objects. Extensive experiments are conducted on four typical visual condition types, and our SmartControl can achieve a superior performance against state-of-the-art methods. Source code, pre-trained models, and datasets will be publicly available.

Keywords: Text-to-Image Generation · ControlNet · Rough Conditions

1 Introduction

Recent advances in diffusion models [10, 28] have made remarkable progress in text-to-image (T2I) generation, and large-scale pre-trained T2I models are capable of generating high-quality images according to given text prompts. Building upon these advances, ControlNet [39] and T2I-Adapter [20] further introduce an extra visual condition (*e.g.*, edge maps, human pose skeletons, segmentation maps, and depth) to pre-trained T2I models for layout-controllable image generation. By combining the text prompts and visual conditions, these models can produce images that match the text prompts and adhere to visual conditions.

Despite the promising results, one of the common challenges is creating suitable layout conditions. In practice, users usually use the conditions extracted from other real images. However, such a condition is usually rough, and cannot provide layout information that is precisely aligned with text prompts during conditional image generation. The inconsistency between the condition and text prompt will lead to the degraded generated results. For example, as shown in the top of Fig. 1, when the user generates the photo of a girl with the visual condition from Mickey, the generated image shows obvious artifacts in the human's head and arm. Nonetheless, handcrafted modifying the visual conditions is a professional and time-consuming task, and even infeasible for users.

To improve the quality of generation on the rough condition, one possible solution is to relax the restriction of visual condition. For example, LooseControl [4] proposes to control the layout of the image through a 3D bounding box, including the position, orientation, and size of the object. Although LooseControl achieves flexible controllability, its visual condition is too loose to control the posture and actions of the objects effectively. Another possible way is to reduce the control intensity of visual conditions. In ControlNet [39], the visual conditions are integrated into the generation process by adding the feature of visual conditions to the latent image features. Therefore, we can decrease the fusion weight of visual conditions to relax its influence, so that the generative models can balance information from text and visual conditions. As shown in Fig. 2, a proper weight may produce a desired result. However, the optimal weights for different inputs are varied, and one should manually navigate all control intensities for selecting a suitable one. Moreover, it is even infeasible to find a suitable weight for some cases (see the third row of Fig. 2). Furthermore, the fusion weight is a global parameter that affects the entire image, leading to compromises between different local areas.

In this work, we propose SmartControl, an automated and flexible method for photo-realistic image generation with the text prompt and a rough visual condition as inputs. We argue that the text prompt can more faithfully reflect user intentions during generation. At the same time, the rough visual conditions usually provide only coarse layout information. Therefore, the key idea of our SmartControl is to relax the constraints on areas that conflict with the text prompts in the rough visual conditions, and we propose a Control Scale Predictor to identify the conflict regions and predict the local control scales based on the visual conditions and text prompts. Considering that both text prompt and conditions are integrated into ControlNet backbone, we extract relevant priors regarding visual conditions and text prompts from the generative model. Then, the control scale predictor can take advantage of the generative priors to predict the spatial adaptive control scales. For training the control scale predictor, a dataset with text prompts and rough visual conditions is constructed. Thanks to the generative prior extracted from the pre-trained generative models, a limited number (*e.g.*, 1,000∼2,000) of samples is sufficient, and our SmartControl shows preferable generalization abilities to unseen objects. As shown in Fig. 1, our SmartControl could generate photo-realistic images faithful to text prompts while preserving useful information from the rough visual conditions.

Extensive experiments are conducted on various backbone generative models and visual condition types, and our SmartControl can perform favorably against state-of-the-art methods. Our contributions are listed as follows:

– We present an automated and flexible text-to-image generation method under rough visual conditions (dubbed SmartControl), which achieves local-adaptive control intensities based on the consistency between text prompts and the visual conditions.

– A control scale predictor is designed to distinguish and identify conflicts between text prompts and visual conditions.

– A dataset with text prompts and unaligned rough visual conditions is constructed, based on which extensive experiments are conducted, showing that our proposed method performs favorably against state-of-the-art methods.

2 Related Work

2.1 Text-to-Image Diffusion Model

Diffusion models [10, 31] have achieved remarkable success in the field of text-to-image (T2I) generation [3, 21, 27, 28, 30], capable of generating images with high fidelity and diversity. T2I diffusion models redefine the image generation task as an iterative denoising process guided on text embeddings produced by language encoders such as CLIP [25] or T5-pretrained [26]. Some methods [3, 27, 30] adopt low-resolution models in pixel domain, coupled with cascaded super-resolution diffusion models. On the other hand, latent diffusion models [28, 36] focus exclusively on performing diffusion processes in the latent space, relying on separately trained high-resolution autoencoders. Stable Diffusion [32] represents a large-scale implementation of the latent diffusion model, which has been widely adopted in various applications, such as controllable image generation [12,16,20, 39], customized image generation [8,29,34], and image manipulation [6,9,19,22].

2.2 Controllable Text-to-Image Generation

Text-to-image diffusion models have achieved promising ability in generating high-fidelity images based on text prompts. However, conveying the desired spatial information solely through text prompts remains a significant challenge. To address this, several approaches have been developed to achieve controllable text-to-image generation by adding conditional control such as pose [5,13], 2D bounding boxes [23], segmentation map [2, 14, 35], and multiple conditions [12, 20, 39] like edge maps, depth maps, segmentation masks, normal maps, and OpenPose.

ControlNet [39] adds visual conditions to a pretrained text-to-image diffusion model through the fine-tuning of trainable encoder copies. T2I-Adapter [20] employs various adapters under different conditions to achieve controllable guidance. Several works have built upon ControlNet [39] to introduce improvements, including mixing modalities [11, 24], efficient architecture [38], and loose control [4]. Cocktail [11] allows for the combination of existing modalities and automatically balances the differences between them. UniControl [24] employs a mixture of expert style adapter and a task-aware HyperNet to unify various Condition-to-Image tasks in a single framework, thus compressing the model size. ControlNet-XS [38] focuses on designing an efficient and effective architecture without information transmission delays. LooseControl [4] presents a novel approach to guiding image generation using 3D box depth conditions, employing generalized guidance to enhance the creative possibilities available to users. However, this approach is overly permissive, focusing only on maintaining position and size, while often neglecting the crucial aspect of pose Unlike previous methods, FreeControl [18] provides a training-free approach for multi-condition T2I generation, enabling structural alignment with guidance images and appearance alignment with images generated without control. In comparison to the aforementioned controllable T2I method, our solution has the ability to handle the rough conditions, thereby ensuring greater flexibility in the image generation.

3 Methodology

3.1 Preliminary

SmartControl takes advantage of pre-trained generative methods, specifically Stable Diffusion [32] and ControlNet [39]. For a clear and comprehensive explanation, we provide a brief introduction to preliminary knowledge and some symbol definitions.

Stable Diffusion [32] is a widely employed text-to-image generation method. Given a text prompt $\mathbf{p}$, Stable Diffusion gradually integrates $\mathbf{p}$ into the image generation process via a text-conditioned cross-attention mechanism. For controlling the layout of the generated images, alongside the pre-trained Stable Diffusion [32], ControlNet [39] further introduces a visual condition $\mathbf{c}$, which can be in the form of edge maps, segmentation masks, and so on. Then the image generation process can be defined by $\mathbf{I} = G(\mathbf{p}, \mathbf{c})$, and the working scheme of ControlNet [39] at layer i of the decoder D can be represented by,

$$\mathbf{h}^{i+1} = D^i(\mathbf{h}^i + \mathbf{h}^i_{cond}), \quad 0 \leq i \leq N-1, \tag{1}$$

where $\mathbf{h}^i$ is the feature in the i-th layer of the Stable Diffusion decoder, and $\mathbf{h}^i_{cond}$ is the feature generated from the visual condition $\mathbf{c}$. In this way, the visual condition is successfully introduced into the generation process, and the images generated by ControlNet [39] are constrained to follow both $\mathbf{p}$ and $\mathbf{c}$.

3.2 Problem Definition

Under the assumption of existing layout-controllable T2I generation methods [11, 12, 20, 24, 39], $\mathbf{p}$ and $\mathbf{c}$ are compatible with each other. However, in practice, preparing a visual condition that precisely aligns with the text prompt and user intentions is difficult or infeasible for ordinary users. Therefore, the visual condition $\mathbf{c}$ is often obtained via cheaper ways, for example, extracting from an existing image. We refer to such $\mathbf{c}$ as rough visual conditions (denoted by $\mathbf{c}_{rough}$) since they are not precisely aligned with the text prompt and the users usually intend to follow these conditions at a coarse scale.

Formally, given such a more practical pair of conditions $(\mathbf{p}, \mathbf{c}_{rough})$, the most intuitive principles should be (i) in the regions of $\mathbf{c}_{rough}$ that aligns with $\mathbf{p}$, the generation degrades to the setting of ControlNet [39], and we can safely follow both conditions, and (ii) for regions that $\mathbf{c}_{rough}$ and $\mathbf{p}$ conflict with each other, we should follow the content of $\mathbf{p}$, and adhere to $\mathbf{c}_{rough}$ as much as possible. In summary, the extent that $\mathbf{c}_{rough}$ and $\mathbf{p}$ conflict determines how much $\mathbf{c}_{rough}$ influences the generation result. Therefore, an intuitive way to achieve the goal of this paper can be represented by rewriting Eq. (1) as,

$$\mathbf{h}^{i+1} = D^i(\mathbf{h}^i + \boldsymbol{\alpha} \cdot \mathbf{h}^i_{cond}), \quad 0 \leq i \leq N-1, \tag{2}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{H \times W}$ denotes a spatial control scale map, which is adaptive to the conflict between $\mathbf{c}_{rough}$ and $\mathbf{p}$. Then, the core task of this work is to design and train a control scale predictor f to predict such a control scale map $\boldsymbol{\alpha} = f(\mathbf{p}, \mathbf{c}_{rough}; \theta)$.

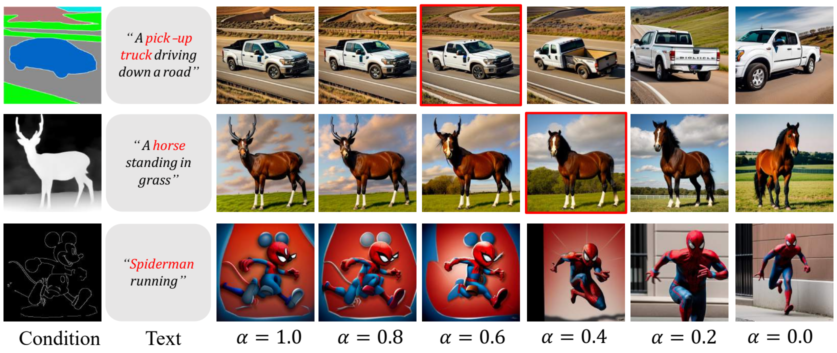

Figure 2: Images generated with different control scale. The plausible images are highlighted in red boxes with different control scale. It is even infeasible to find a suitable control scale for some cases.

3.3 Control Scale Exploration

For training the control scale predictor f and achieving the adaptive control scale in Eq. (2), most challenging issue is the lack of supervision information for $\boldsymbol{\alpha}$. In order to explore the rules of the control scale and design reasonable training criteria for training f, we first degrade $\boldsymbol{\alpha}$ to a scalar α that denotes the global control scale for the whole image, and tune the value of α to show its influence.

As shown in Fig. 2, when given a pair of an unaligned text prompt $\mathbf{p}$ and a rough visual condition $\mathbf{c}_{rough}$, ControlNet [39] (*i.e.*, when $\alpha = 1.0$) strictly follows the layout of $\mathbf{c}_{rough}$ and fits the object mentioned in the text prompt $\mathbf{p}$ into the shape described by $\mathbf{c}_{rough}$. For example, a deer antler is added to the horse, and two round ears appear on the head of Spider-Man. By gradually decreasing the value of α, one can see that the generated images become better aligned with the text prompt $\mathbf{p}$, until the effect of $\mathbf{c}_{rough}$ disappears when $\alpha = 0.0$.

Besides, we have also observed a large amount of samples, drawing the following conclusions. (i) For a portion of the $(\mathbf{p}, \mathbf{c}_{rough})$ pairs, a proper control scale α can be found to generate a plausible image[4]. (ii) Even if the optimal α is not found, it seems promising to obtain a desired image by combining results with different α. For example, as shown in the third row of Fig. 2, we can get a potential result by combining Spider-Man when $\alpha = 0.6$ and background when $\alpha = 0.4$. (iii) For areas that $\mathbf{c}_{rough}$ conflicts with $\mathbf{p}$, a large enough freedom (*i.e.*, small enough α) should be assigned to breaking free from the constraints of $\mathbf{c}_{rough}$. On the contrary, a sufficiently large α should be set in other areas to ensure the effectiveness of $\mathbf{c}_{rough}$. Among these three observations, the second item shows that the proposed method in Eq. (2) is feasible, while the first and third ones provide the possibility to construct the dataset and train the control

[4] For the $(\mathbf{p}, \mathbf{c}_{rough})$ pairs we delicately prepared in Sec. 3.5, we can find a suitable α for around 60%∼70% of the samples.

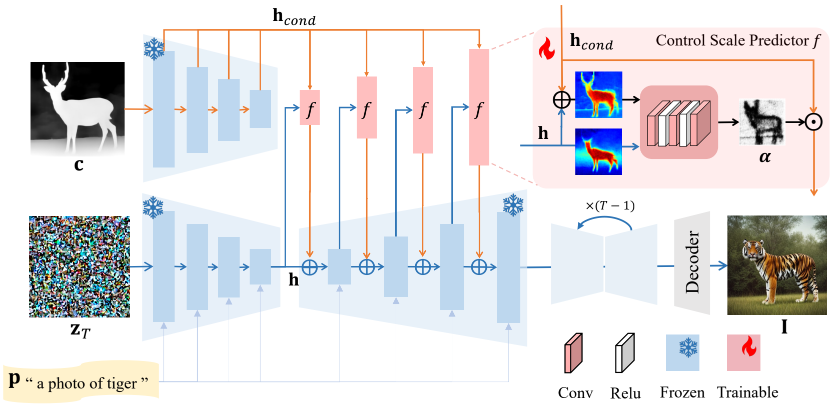

Figure 3: Framework of proposed SmartControl. Our method is built upon ControlNet, and can generate photo-realistic images with inconsistent prompt and rough visual condition (*i.e.*, tiger *v.s.* deer) as input. To achieve this, we introduce a control scale predictor f for each decoder block of ControlNet. The predictor takes $\mathbf{h}$ and $\mathbf{h} + \mathbf{h}_{cond}$ as input and predicts a pixel-wise control scale map $\boldsymbol{\alpha}$. The condition feature $\mathbf{h}_{cond}$ is then updated by $\boldsymbol{\alpha} \cdot \mathbf{h}_{cond}$ to relax the control scale at conflict region, resulting a plausible and photo-realistic generated image.

scale predictor f. In the following, we will introduce the implementation of the control scale predictor and show the pipeline to construct the dataset in detail.

3.4 Control Scale Predictor

Generative Prior Extraction. Among the above analysis, f needs to identify visual concept of the object described in the text prompts and locate the inconsistency between it and visual conditions. This requires our predictor to have the ability to fully comprehend the prompts and visual conditions, and such a network typically requires a substantial dataset for training. To reduce the training requirements, we propose leveraging the superior capabilities of ControlNet as a prior. Building upon Stable Diffusion [28], ControlNet can extract meaningful features from prompts and visual conditions separately and utilize them to generate desired images. Specifically, $\mathbf{h} + \mathbf{h}_{cond}$ integrates information about the visual condition, while $\mathbf{h}$ encodes the information from the given prompt. Therefore, instead of using $\mathbf{p}$ and $\mathbf{c}$ as inputs to the predictor, we utilize $\mathbf{h}$ and $\mathbf{h} + \mathbf{h}_{cond}$, which facilitates the easier identification of inconsistencies.

Network Architecture. The overall architecture of proposed SmartControl is illustrated in Fig. 3. Within each decoder block D^i, we incorporate a control scale predictor f^i to predict spatially adaptive control scales $\boldsymbol{\alpha}^i$. The control scale predictor consists of three stacked modules (each containing a convolutional layer and a ReLU layer) and a sigmoid function. The i-th predictor takes $\mathbf{h}^i$ and

$\mathbf{h}^i + \mathbf{h}^i_{cond}$ as input and predicts a pixel-wise control scale map $\boldsymbol{\alpha}^i$,

$$\boldsymbol{\alpha}^i = f^i(\mathbf{h}^i, \mathbf{h}^i + \mathbf{h}^i_{cond}), \quad 0 \leq i \leq N-1. \tag{3}$$

As depicted in Fig. 3, the predicted $\boldsymbol{\alpha}$ exhibits minimal values in the conflict region (antlers and legs), while approaching 1.0 in other regions (background). This indicates that the predicted $\boldsymbol{\alpha}$ is plausible, and we can utilize $\boldsymbol{\alpha}$ to generate the desired image of a tiger.

Learning Objective. Following ControlNet [39], we employ the mean-squared loss to train our predictor,

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathbf{z}_0,\mathbf{t},\mathbf{p},\mathbf{c}_{rough},\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{t}, \mathbf{p}, \mathbf{c}_{rough}))\|_2^2], \tag{4}$$

where ϵ_θ denotes our model and $\mathbf{z}_0$ represents the latent embedding of real image. ϵ denotes the unscaled noise and $\mathbf{t}$ denotes the time step of diffusion process. $\mathbf{z}_t$ is the latent noise at $\mathbf{t}$ step.

To provide explicit supervision for the control scale predictor f, we further introduce a regularization term to ensure that the values of $\boldsymbol{\alpha}$ should be maintained above α_{bg} in the background regions and below $\alpha_{conflict}$ in the conflict regions,

$$\mathcal{L}_{\text{c}} = \mathbf{m}_{conflict} \cdot \max(\mathbf{0}, \boldsymbol{\alpha} - \alpha_{conflict}) + \mathbf{m}_{bg} \cdot \max(\mathbf{0}, \alpha_{bg} - \boldsymbol{\alpha}), \tag{5}$$

where α_{bg} and $\alpha_{conflict}$ are hyper-parameters. $\mathbf{m}_{conflict}$ denotes the mask of conflict areas, and $\mathbf{m}_{bg}$ is the mask of background.

The overall learning objective for training the SmartControl is defined by,

$$\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda_{\text{c}}\mathcal{L}_{\text{c}}, \tag{6}$$

where λ_{c} is hyper-parameters for balancing different loss terms.

3.5 Unaligned Data Construction Pipeline

ControlNet [39] utilizes an image $\mathbf{I}$ as input and generates aligned conditions $\mathbf{c}$ and text prompt $\mathbf{p}$, which is not suitable to train our SmartControl. In this section, we will introduce the workflow for constructing the unaligned text-condition dataset as shown in Fig. 4. Specifically, we first generate the unaligned visual conditions and text. Then, the paired image is generated by ControlNet [39] based on these unaligned texts and conditions.

Generating Unaligned Visual Conditions and Text Prompts. The original image is from OpenImage [15] and contains an object occupying over 30% of the image area. The visual condition is generated by a pre-existing condition estimator. To create a plausible inconsistent text prompt for the given image with class <$\texttt{cls}_{init}$>, we first employ the hierarchical class tree to determine one alternative class <$\texttt{cls}_{alt}$>, where <$\texttt{cls}_{init}$> and <$\texttt{cls}_{alt}$> share the same parent class. Then, the inconsistent text prompt can be formatted as "`a photo of a <cls`$_{alt}$`>`.". For example, as shown in Fig. 4, if the rough depth condition

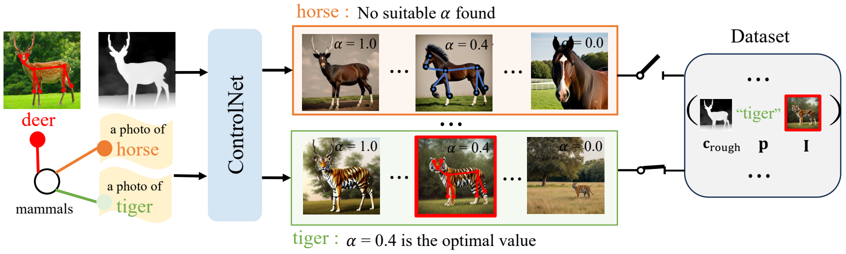

Figure 4: Pipeline for unaligned data construction. Given an image and corresponding class, we extract the visual condition $\mathbf{c}_{rough}$ (*e.g.*, depth) by the pre-trained estimator. Then, for the given class (*e.g.*, deer), we select an alternative unaligned class (*e.g.*, tiger or horse) based on class hierarchy, and use it to obtain the unaligned prompt $\mathbf{p}$. By iterating through different control scale α of ControlNet [39], we can generate a series of images for $(\mathbf{c}_{rough}, \mathbf{p})$. Then, we manually filter those images that are faithful to both text and rough condition to construct our dataset. For example, for tiger, the image generated with $\alpha = 0.4$ is plausible and is added to our dataset. While for horse, there is not a suitable image and all images are discarded.

$\mathbf{c}_{rough}$ is from an image of a deer, the corresponding prompt $\mathbf{p}$ is "`a photo of a horse.`".

Generating Paired Images Based on ControlNet [39]. From the above analysis, we iterate over different control scale α, *i.e.*, $\alpha \in \{1.0, 0.8, 0.6, 0.4, 0.2, 0.0\}$, to perform sampling based on ControlNet [39], followed by manual filtering. If a proper control scale α can be found to generate a plausible image, we add this data into our dataset. In cases where no proper α is found, the data is discarded. It is worth noting that our iteration is limited within a dilation range of the chosen object and the value of α is 1.0 in other areas. Besides, we also acquire $\mathbf{m}_{conflict}$ and $\mathbf{m}_{bg}$ for training. $\mathbf{m}_{conflict}$ denotes the areas of conflict between $\mathbf{c}_{rough}$ and $\mathbf{p}$, *i.e.*, different part between the mask of <cls$_{init}$> and <cls$_{alt}$>, while $\mathbf{m}_{bg}$ represents the background region, defined as follows,

$$\mathbf{m}_{conflict} = |\mathbf{m}_{alt} - \mathbf{m}_{init}|, \mathbf{m}_{bg} = 1 - (\mathbf{m}_{alt} \vee \mathbf{m}_{init}), \tag{7}$$

where $\mathbf{m}_{alt}$ and $\mathbf{m}_{init}$ are obtained based on the existing segmentation method [40] based on "`a photo of a <cls`$_{alt}$`>.`" and "`a photo of a <cls`$_{init}$`>.`".

4 Experiments

4.1 Experimental Details

Datasets. We collect training datasets across four types of conditions including depth, HED, segmentation, and canny. The dataset sizes for each condition are 2,000, 1,500, 1,500, and 1,000 images respectively. For each condition type, we collect an evaluation dataset of 100 images including 70 images with significant conflicts, 20 images with mild conflicts, and 10 conflict-free images to assess the performance in handling diverse conditions. Our evaluation dataset includes 48 classes, and 12.5% of those classes do not appear in the training dataset, which allows us to evaluate the generalization ability.

Evaluation Metrics. Following [33], we use *CLIP Score* [25] metric to measure text-image alignment and use the *Self-similarity distance* metric to measure the structural similarity between two images in the feature space of the DINO-ViT model [7]. A smaller Self-similarity distance implies that the generated image closely preserves the structure of the original image, Moreover, we introduce a metric named *Class Confidence* to assess whether the generated images align with the desired class. A higher *Class Confidence* indicates that the generated images closely match the desired class, not affected by the inherent class of the rough conditions. To comprehensively evaluate structure preservation and image-text alignment, we propose to utilize GPT-4V [1] as a novel metric. Given two images from different methods, we ask GPT-4V [1] to determine which of them is better by examining through two aspects: first, whether the pose or layout matches the reference image, and second, whether it aligns more accurately with the given text. The specific prompt can be found in the supplementary material.

Implementation Details. In all our experiments, we train our control scale predictor based on the pre-trained ControlNet [39], while keeping all parameters of ControlNet [39] fixed. The model is trained with an AdamW [17] optimizer with weight decay of 1×10^{-5} for 200 epochs. The trade-off parameter λ_c is determined to be 0.01. Furthermore, $\alpha_{conflict}$ and α_{bg} are set at 0.2 and 0.8 respectively.

4.2 Comparison with Existing Methods

We choose the following state-of-the-art models in controllable image generation as competing methods: ControlNet [39], T2I-Adapter [20] and Uni-ControlNet [24]. However, standard ControlNet [39] and T2I-Adapter [20] are not suitable for rough conditions. For a fair comparison, we employed the small control scale α_{fix}, for both ControlNet [39] and T2I-Adapter [20]. Here, α_{fix} represents the optimal but fixed control scale for the entire evaluation dataset. However, α_{fix} varies across different modalities. For example, we use $\alpha_{fix} = 0.4$ for the depth conditions and $\alpha_{fix} = 0.6$ for the segmentation conditions in ControlNet [39].

Quantitative Comparison. We conduct comprehensive experiments in four types of conditions to assess the effectiveness of the proposed method, and the quantitative results are shown in Tab. 1. We can observe that while ControlNet (α=1.0) [39] and T2I-Adapter($\alpha = 1.0$) [20] are stronger in maintaining structure, they struggle to generate images aligned with textual prompts, resulting in significantly lower CLIP Scores. ControlNet($\alpha = \alpha_{fix}$) [39] and T2I-Adapter($\alpha = \alpha_{fix}$) [20] show inferior performance as they exhibit limitations in handling diverse text prompts and structural conditions. Our method instead exhibits significant improvement in CLIP Scores compared to the previous methods, indicating improved structural similarity and image-text alignment. Note that We did not achieve a superior Self-similarity metric in the unpaired evaluation dataset. However, a low Self-similarity metric may indicate that the

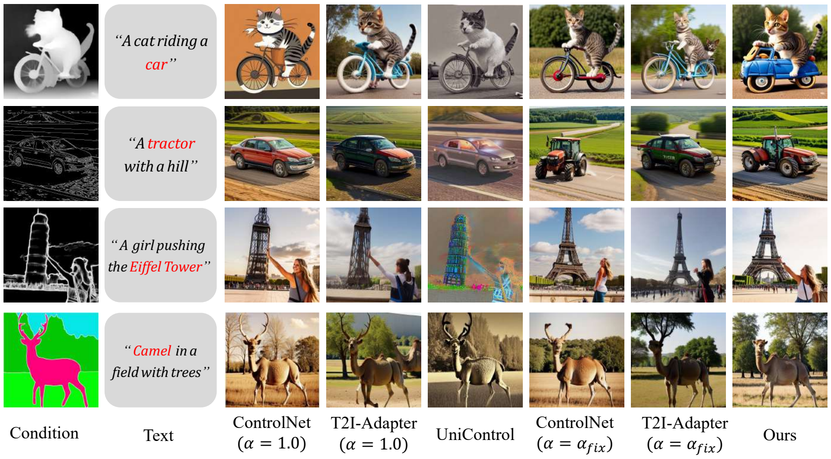
"A cat riding a car"
"A tractor with a hill"
"A girl pushing the Eiffel Tower"
"Camel in a field with trees"
Condition
Text
ControlNet (α = 1.0)
T2I-Adapter (α = 1.0)
UniControl
ControlNet ($\alpha = \alpha_{fix}$)
T2I-Adapter ($\alpha = \alpha_{fix}$)
Ours

Figure 5: Qualitative comparison with different modalities, image prompts and additional visual conditions. SmartControl achieves reasonable spatial control and superior image-text alignment compared to existing methods, resulting in a closer match to human intentions.

Table 1: Quantitative comparison for Controllable Text-to-Image Generation for rough conditions on our evaluation dataset.The best results are highlighted with **bold**.

Method	Depth			Canny			Seg			HED		
	CLIP↑	CLASS↑	Sim↓	CLIP↑	CLASS↑	Sim↓	CLIP↑	CLASS↑	Sim↓	CLIP↑	CLASS↑	Sim↓
ControlNet(α=1.0) [39]	0.257	0.602	**0.100**	0.244	0.467	**0.107**	0.258	0.666	**0.115**	0.264	0.647	0.123
T2I Adapter(α=1.0) [20]	0.267	0.593	0.123	0.253	0.464	0.109	0.251	0.492	0.138	0.261	0.621	0.106
UniContrtol [24]	0.251	0.597	0.102	0.240	0.379	0.117	0.261	0.668	0.116	0.227	0.336	**0.082**
ControlNet($\alpha = \alpha_{fix}$) [39]	0.268	0.710	0.136	0.270	**0.736**	0.149	0.267	0.696	0.140	0.271	0.727	0.143
T2I Adapter($\alpha = \alpha_{fix}$) [20]	0.271	0.721	0.137	**0.272**	0.682	0.141	0.263	0.668	0.143	0.269	0.747	0.137
Ours	**0.274**	**0.742**	0.128	**0.272**	0.721	0.143	**0.277**	**0.780**	0.140	**0.276**	**0.768**	0.142

generated images overly adhere to the rough conditions, which does not always equate to better performance for rough conditions. In the supplementary material, we will provide a Self-similarity metric calculated with pseudo-ground truths instead of the original images. Considering the significant effort for utilizing GPT-4V [1] as the metric, we select the commonly used condition, *i.e.*, depth to compare our method with ControlNet($\alpha = \alpha_{fix}$) [39]. In the majority of cases, specifically 67%, GPT-4V [1] ranked our result better.

Qualitative Comparison. The qualitative results of competing methods are shown in Fig. 5. ControlNet (α = 1.0) [39], T2I-Adapter (α = 1.0) [20], and UniControl, when constrained by rough conditions, generate images that are unrealistic and misaligned with the text prompts. Meanwhile, ControlNet ($\alpha = \alpha_{fix}$) [39] and T2I-Adapter ($\alpha = \alpha_{fix}$) [20] offer some improvement in specific scenarios. However, due to the global and uniform α across all images, they still encounter failure in numerous situations. In the example from the first row, **"**`a cat is driving a car.`**"**, there is a conflict in `car`. Despite altering the pose of cat, it is not possible to successfully transform from a bicycle to a car

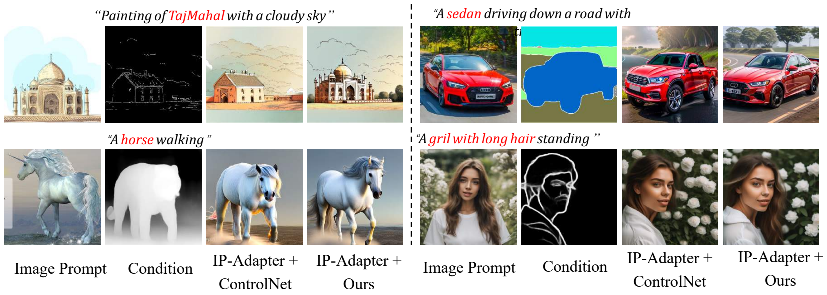

Figure 6: Visualization of generated samples with the IP-Adapterr [37]. Note that we do not need fine-tune our control scale predictor.

using ControlNet ($\alpha = \alpha_{fix}$) [39] and T2I-Adapter ($\alpha = \alpha_{fix}$) [20]. Moreover, in cases where it is necessary to remove regions (which require extremely small values of α), such as in the fourth row example, all competing methods would result in images retaining the deer antlers. As illustrated in Fig. 5, our proposed SmartControl is capable of generating images that not only closely resemble real images but also align more accurately with text prompts and useful information in rough conditions, demonstrating its superior performance. More qualitative results are given in the supplementary material.

User Study. We invite 20 users to participate in our user study to assess the effectiveness of our methods. We utilize six different methods respectively and generate 40 images for each method based on different types of visual conditions and text prompts. Each user is requested to select the best image based on the text-image alignment and structural similarity with visual conditions. In the majority of cases, *i.e.*, 78.3%, users prefer our method.

4.3 More Results

While originally designed for rough conditions, SmartControl demonstrates robust generalization capabilities, enabling it to effortlessly adapt to other models without retraining. In this section, we showcase additional results through the integration of our SmartControl with the IP-Adapter [37]. The primary function of IP-Adapter [37] is to interpret image prompts to pre-trained text-to-image diffusion models. Fig. 6 shows that the images generated by SmartControl are not only more captivating but also more coherent with image prompt under the rough conditions.

4.4 Ablation Study

Effect of Training Dataset Sizes. Even with a limited set of 0.5k images, our training process remains stable, and as the dataset size increases, the realism of the generated images improves (as shown in Fig. 7 and Tab. 2). Obviously, we choose 2k images for our training dataset under the depth condition. Although

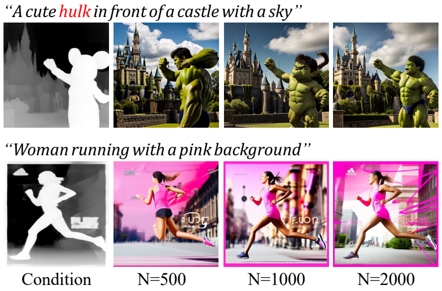
"A cute hulk in front of a castle with a sky"
"Woman running with a pink background"
Condition
N=500
N=1000
N=2000

Figure 7: Visualisation of ablation study different training dataset sizes.

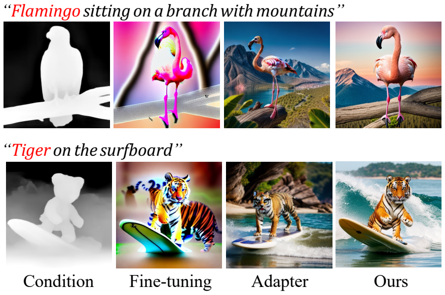

Figure 8: Visual comparison for different implementation methods based on our dataset.

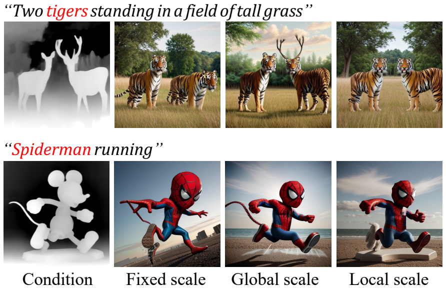

Figure 9: Visual comparison for the different granularity of control scale.

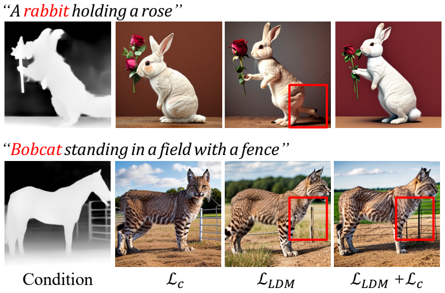
"A rabbit holding a rose"
"Bobcat standing in a field with a fence"
Condition
$\mathcal{L}_c$
$\mathcal{L}_{LDM}$
$\mathcal{L}_{LDM}+\mathcal{L}_c$

Figure 10: Impact of $\mathcal{L}_{\text{LDM}}$ and $\mathcal{L}_{\text{c}}$ contribute to the overall performance.

Table 2: Ablation study on the sizes of training dataset under the depth condition.

Datasets	CLIP↑	CLASS↑	Sim↓
N=500	0.273	0.740	0.134
N=1000	**0.274**	0.724	0.130
N=2000	**0.274**	**0.742**	**0.128**

Table 3: Effect of the proposed control scale predictor for rough conditions.

Method	CLIP↑	CLASS↑	Sim↓
Fine-tuning	0.248	0.307	0.137
Adapter	0.273	0.731	0.192
Ours	**0.274**	**0.742**	**0.128**

Table 4: Effect of local control scale.

Method	CLIP↑	CLASS↑	Sim↓
Fixed Scale α_{fix}	0.268	0.710	0.136
Global Scale α_{glob}	0.272	0.741	**0.122**
Local Scale $\boldsymbol{\alpha}$	**0.274**	**0.742**	0.128

Table 5: Ablation of the network architecture for the control scale predictor.

Method	CLIP↑	CLASS↑	Sim↓	Time↓
Cross Atten	0.272	0.734	**0.126**	7.69
Conv(Ours)	**0.274**	**0.742**	0.128	**7.36**

the dataset consists of only 2k images, we achieve commendable results across the open domain. More analysis and visual results of generalization capability are provided in the supplementary material.

Effect of Control Scale Predictor. As illustrated in Sec. 3.4, we apply a control scale predictor to adaptively adjust the control intensity based on various conditions, and text prompts. In this subsection, we make detailed experiments to assess the effects of control scale predictor, *e.g.*, the fine-tuning scheme, the granularity of the control scale, and the network structure.

Fine tuning scheme. In order to assess the effect on the control scale predictor, we experiment on several commonly used fine-tuning schemes, *e.g.*, fine-tuning the ControlNet branch, an adapter, and a control scale predictor. Fine-tuning the ControlNet may suffer from the degradation of generation capability. This is evident from a performance drop in CLIP Score and the poor quality of the generated images. On the other hand, training an adapter on our dataset may lead to overfitting, resulting in decreased structural alignment during testing as shown in Fig. 8. As shown in Tab. 3, although CLIP Scores are comparable, our method achieves a 33.3% improvement on Self-similarity metric over the adapter, which demonstrates the effectiveness of the control scale predictor.

Granularity of control scale. In this subsection, we make detailed experiments to assess the effect of different granularity of control scale, *e.g.*, fixed scale a_{fix}, global scale a_{glob} and local scale map $\boldsymbol{\alpha}$. Using a fixed scale that is applied uniformly across the entire evaluation dataset leads to a decrease in performance and the generation of lower-quality images. We trained our model to predict the global scale based on our dataset. However, the global scale is insufficient to handle situations where the required control scale varies within a single image. For example, in the second line in Fig. 9, the tail is not thoroughly removed, and the base part is not preserved. In contrast, our method is designed to predict the local control scale, which effectively addresses these issues. As shown in Tab. 4, the performance is promoted with local control scale, which also demonstrates the effectiveness of the local control scale.

Network architecture. We experiment with commonly used network architectures, *e.g.*, convolution, and cross-attention. The experimental results in Tab. 5 revealed that all of them yield better performance. We selected the convolution to implement the control scale predictor in this paper as it is relatively faster.

Effects of Loss. We investigate the impact of $\mathcal{L}_{\mathrm{LDM}}$ and $\mathcal{L}_{\mathrm{c}}$ in Fig. 10. The model trained solely with $\mathcal{L}_{\mathrm{LDM}}$ exhibits relatively poor performance and lacks accuracy in constraining the layout. Training without $\mathcal{L}_{\mathrm{c}}$ leads to the reduction of control even in non-conflicting areas, such as the background in the second row. Additionally, it results in residual artifacts such as the tail of the rabbit.

5 Conclusion

In this paper, we introduce a SmartControl, a flexible controllable image generation under rough visual conditions. Unlike existing approaches, SmartControl adaptively handles situations where there are inconsistencies between visual conditions and text prompts. We introduce the control scale predictor, capable of identifying displacement regions between visual conditions and prompts and predicting local adaptive control strengths based on the displaces. For training and evaluation, we construct a dataset with unaligned text prompts and visual conditions. Extensive experiments demonstrate that our SmartControl achieves better performance against the state-of-the-art methods under rough visual conditions.

Acknowledgement

This work was supported in part by the National Key R&D Program of China under Grant No. 2022YFA1004100 and the National Natural Science Foundation of China (NSFC) under Grant No. 62441202.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18370–18380 (2023)
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Bhat, S.F., Mitra, N.J., Wonka, P.: Loosecontrol: Lifting controlnet for generalized depth conditioning. arXiv preprint arXiv:2312.03079 (2023)
5. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Laaksonen, J., Shah, M., Khan, F.S.: Person image synthesis via denoising diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5968–5976 (2023)
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
8. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
9. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Hu, M., Zheng, J., Liu, D., Zheng, C., Wang, C., Tao, D., Cham, T.J.: Cocktail: Mixing multi-modality controls for text-conditional image generation. arXiv preprint arXiv:2306.00964 (2023)
12. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
13. Ju, X., Zeng, A., Zhao, C., Wang, J., Zhang, L., Xu, Q.: Humansd: A native skeleton-guided diffusion model for human image generation. arXiv preprint arXiv:2304.04269 (2023)
14. Kim, Y., Lee, J., Kim, J.H., Ha, J.W., Zhu, J.Y.: Dense text-to-image generation with attention modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7701–7711 (2023)

15. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
16. Li, X., Hou, X., Loy, C.C.: When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2187–2196 (2024)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Mo, S., Mu, F., Lin, K.H., Liu, Y., Guan, B., Li, Y., Zhou, B.: Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. arXiv preprint arXiv:2312.07536 (2023)
19. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
20. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
21. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
22. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
23. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. arXiv preprint arXiv:2306.05427 (2023)
24. Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J.C., Xiong, C., Savarese, S., et al.: Unicontrol: A unified diffusion model for controllable visual generation in the wild. arXiv preprint arXiv:2305.11147 (2023)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
26. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
27. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
29. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023)
30. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

31. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
32. Stability: Stable diffusion v1.5 model card (2022), `https://huggingface.co/runwayml/stable-diffusion-v1-5`
33. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plugand-play diffusion features for text-driven image-toimage translation. arXiv preprint arXiv:2211.12572 (2022)
34. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
35. Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W.: Freestyle layout-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14256–14266 (2023)
36. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. arXiv preprint arXiv:2305.18295 (2023)
37. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
38. Zavadski, D., Feiden, J.F., Rother, C.: Controlnet-xs: Designing an efficient and effective architecture for controlling text-to-image diffusion models. arXiv preprint arXiv:2312.06573 (2023)
39. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
40. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024)