FlexEControl: Flexible and Efficient Multimodal Control for Text-to-Image Generation

Anonymous authors
Paper under double-blind review

Abstract

Controllable text-to-image (T2I) diffusion models generate images conditioned on both text prompts and semantic inputs of other modalities like edge maps. Nevertheless, current controllable T2I methods commonly face challenges related to efficiency and faithfulness, especially when conditioning on multiple inputs from either the same or diverse modalities. In this paper, we propose a novel *Flexible* and *Efficient* method, FlexEControl, for controllable T2I generation. At the core of FlexEControl is a unique weight decomposition strategy, which allows for streamlined integration of various input types. This approach not only enhances the faithfulness of the generated image to the control, but also significantly reduces the computational overhead typically associated with multimodal conditioning. Our approach achieves a reduction of 41% in trainable parameters and 30% in memory usage compared with Uni-ControlNet. Moreover, it doubles data efficiency and can flexibly generate images under the guidance of multiple input conditions of various modalities.

1 Introduction

In the realm of text-to-image (T2I) generation, diffusion models exhibit exceptional performance in transforming textual descriptions into visually accurate images. Such models exhibit extraordinary potential across a plethora of applications, spanning from content creation (Rombach et al., 2022; Saharia et al., 2022b; Nichol et al., 2021; Ramesh et al., 2021a; Yu et al., 2022; Avrahami et al., 2023; Chang et al., 2023), image editing (Balaji et al., 2022; Kawar et al., 2023; Couairon et al., 2022; Zhang et al., 2023; Valevski et al., 2022; Nichol et al., 2021; Hertz et al., 2022; Brooks et al., 2023; Mokady et al., 2023), and also fashion design (Cao et al., 2023). We propose a new unified method that can tackle two problems in text-to-image generation: improve the training efficiency of T2I models concerning memory usage, computational requirements, and a thirst for extensive datasets (Saharia et al., 2022a; Rombach et al., 2022; Ramesh et al., 2021b); and improve their controllability especially when dealing with multimodal conditioning, e.g. multiple edge maps and at the same time follow the guidance of text prompts, as shown in Figure 1 (c).

Controllable text-to-image generation models (Mou et al., 2023) often come at a significant training computational cost, with linear growth in cost and size when training with different conditions. Our approach can improve the training efficiency of existing text-to-image diffusion models and unify and flexibly handle different structural input conditions all together. We take cues from the efficient parameterization strategies prevalent in the NLP domain (Pham et al., 2018; Hu et al., 2021; Zaken et al., 2021; Houlsby et al., 2019) and computer vision literature (He et al., 2022). The key idea is to learn shared decomposed weights for varied input conditions, ensuring their intrinsic characteristics are conserved. Our method has several benefits: It not only achieves greater compactness (Rombach et al., 2022), but also retains the full representation capacity to handle various input conditions of various modalities; Sharing weights across different conditions contributes to the data efficiency; The streamlined parameter space aids in mitigating overfitting to singular conditions, thereby reinforcing the flexible control aspect of our model.

Meanwhile, generating images from multiple homogeneous conditional inputs, especially when they present conflicting conditions or need to align with specific text prompts, is challenging. To further augment our model's capability to handle multiple inputs from either the same or diverse modalities as shown in Figure 1,

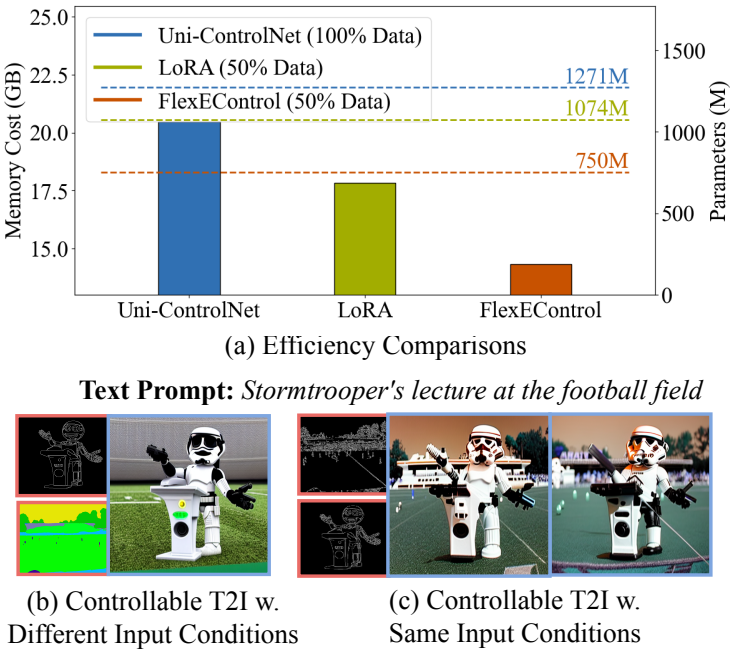

Figure 1: (a) FlexEControl excels in training efficiency, achieving superior performance with just half the training data compared to its counterparts on (b) Controllable Text-to-Image Generation w. Different Input Conditions (one edge map and one segmentation map). (c) FlexEControl effectively conditions on two canny edge maps. The text prompt is `Stormtrooper's lecture at the football field` in both Figure (b) and Figure (c).

during training, we introduce a new training strategy with two new loss functions introduced to strengthen the guidance of corresponding conditions. This approach, combined with our compact parameter optimization space, empowers the model to learn and manage multiple controls efficiently, even within the same category (e.g., handling two distinct segmentation maps and two separate edge maps). Our primary contributions are summarized below:

- We propose FlexEControl, a novel text-to-image generation model for efficient controllable image generation that substantially reduces training memory overhead and model parameters through decomposition of weights shared across different conditions.
- We introduce a new training strategy to improve the flexible controllability of FlexEControl. Compared with previous works, FlexEControl can generate new images conditioning on multiple inputs from diverse compositions of multiple modalities.
- FlexEControl shows on-par performance with Uni-ControlNet (Zhao et al., 2023) on controllable text-to-image generation with 41% less trainable parameters and 30% less training memory. Furthermore, FlexEControl exhibits enhanced data efficiency, effectively doubling the performance achieved with only half amount of training data.

2 Method

The overview of our method is shown in Figure 2. In general, we use the copied Stable Diffusion encoder (Stable Diffusion encoder block and Stable Diffusion middle block) which accepts structural conditional input

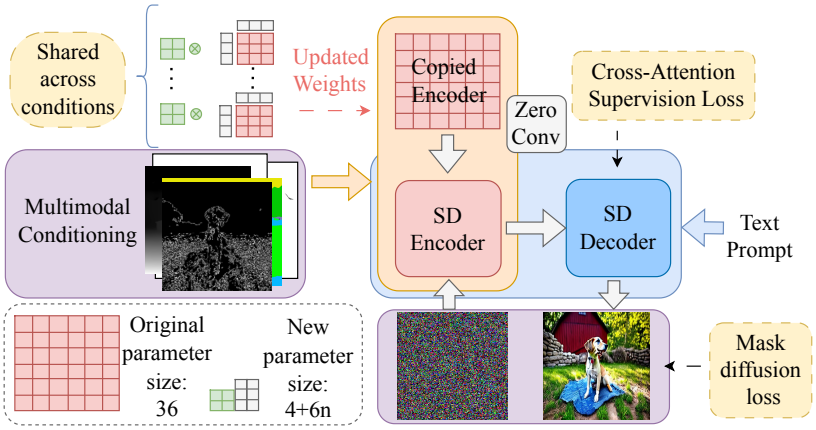

Figure 2: Overview of FlexEControl: a decomposed green matrix is shared across **different input conditions**, significantly enhancing the model's efficiency and preserving the image content. During training, we integrate two specialized loss functions to enable flexible control and to adeptly manage conflicting conditions. In the example depicted here, the new parameter size is efficiently condensed to $4 + 6n$, where n denotes the number of decomposed matrix pairs.

and then perform efficient training via parameter reduction using Kronecker Decomposition first (Zhang et al., 2021a) and then low-rank decomposition over the updated weights of the copied Stable Diffusion encoder. To enhance the control from language and different input conditions, we propose a new training strategy with two newly designed loss functions. The details are shown in the sequel.

2.1 Preliminary

We use Stable Diffusion 1.5 (Rombach et al., 2022) in our experiments. This model falls under the category of Latent Diffusion Models (LDM) that encode input images x into a latent representation z via an encoder $\mathcal{E}$, such that $z = \mathcal{E}(x)$, and subsequently carry out the denoising process within the latent space $\mathcal{Z}$. An LDM is trained with a denoising objective as follows:

$$\mathcal{L}_{\text{ldm}} = \mathbb{E}_{z,c,e,t}\left[\|\hat{\epsilon}_\theta(z_t \mid c, t) - \epsilon\|^2\right] \tag{1}$$

where (z, c) constitute data-conditioning pairs (comprising image latents and text embeddings), $\epsilon \sim \mathcal{N}(0, I)$, $t \sim \text{Uniform}(1, T)$, and θ denotes the model parameters.

2.2 Efficient Training for Controllable Text-to-Image (T2I) Generation

Our approach is motivated by empirical evidence that Kronecker Decomposition (Zhang et al., 2021a) effectively preserves critical weight information. We employ this technique to encapsulate the shared relational structures among different input conditions. Our hypothesis posits that by amalgamating diverse conditions with a common set of weights, data utilization can be optimized and training efficiency can be improved. We focus on decomposing and fine-tuning only the cross-attention weight matrices within the U-Net (Ronneberger et al., 2015) of the diffusion model, where recent works (Kumari et al., 2023) show their dominance when customizing the diffusion model. As depicted in Figure 2, the copied encoder from the Stable Diffusion will accept conditional input from different modalities. During training, we posit that these modalities, being transformations of the same underlying image, share common information. Consequently, we hypothesize that the updated weights of this copied encoder, $\Delta \boldsymbol{W}$, can be efficiently adapted within a shared decomposed low-rank subspace. This leads to:

$$\Delta \boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{H_i} \otimes \left(u_i v_i^\top\right) \tag{2}$$

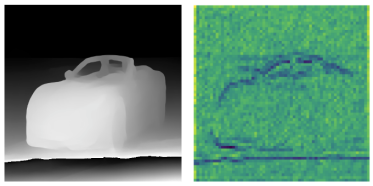

Figure 3: The visualization of decomposed shared "slow" weights (right image) for single condition case where the input condition (left image) is the depth map and the input text prompt is `Car`. We took the average over the decomposed shared weights of the last cross-attention block across all attention heads in Stable Diffusion. The illustrated results shows the content-preserving properties of Kronecker Decompsoition in the task.

with n is the number of decomposed matrices, $u_i \in \mathbb{R}^{\frac{k}{n} \times r}$ and $v_i \in \mathbb{R}^{r \times \frac{d}{n}}$, where r is the rank of the matrix which is a small number, $\boldsymbol{H_i}$ are the decomposed learnable matrices shared across different conditions, and $\otimes$ is the Kronecker product operation. The low-rank decomposition ensures a consistent low-rank representation strategy. This approach substantially saves trainable parameters, allowing efficient fine-tuning over the downstream text-to-image generation tasks.

The intuition for why Kronecker decomposition works for finetuning partially is partly rooted in the findings of Zhang et al. (2021a); Mahabadi et al. (2021); He et al. (2022). These studies highlight how the model weights can be broken down into a series of matrix products and thereby save parameter space. As shown in Figure 2, the original weights is 6x6, then decomposed into a series of matrix products. When adapting the training approach based on the decomposition to controllable T2I, the key lies in the shared weights, which, while being common across various conditions, retain most semantic information. The Kronecker Decomposition is known for its multiplicative rank property and content-preserving qualities. For instance, the shared "slow" weights (Wen et al., 2020) of an image, combined with another set of "fast" low-rank weights, can preserve the original image's distribution without a loss in semantic integrity, as illustrated in Figure 3. This observation implies that updating the slow weights is crucial for adapting to diverse conditions. Following this insight, it becomes logical to learn a set of condition-shared decomposed weights in each layer, ensuring that these weights remain consistent across different scenarios. The data utilization and parameter efficiency is also improved.

2.3 Enhanced Training for Conditional Inputs

We then discuss how to improve the control under multiple input conditions of varying modalities with the efficient training approach.

Dataset Augmentation with Text Parsing and Segmentation To optimize the model for scenarios involving multiple homogeneous (same-type) conditional inputs, we initially augment our dataset. We utilize a large language model (`gpt-3.5-turbo`) to parse texts in prompts containing multiple object entities. The parsing query is structured as: `Given a sentence, analyze the objects in this sentence, give me the objects if there are multiple.` Following this, we apply CLIPSeg (Lüddecke and Ecker, 2022) (`clipseg-rd64-refined` version) to segment corresponding regions in the images, allowing us to divide structural conditions into separate sub-feature maps tailored to the parsed objects. These segmentation masks are selectively used to augment the dataset, specifically when there is a clear, single mask for each identified object. This selective approach helps maintain the robustness of the dataset and enhances training performance.

Cross-Attention Supervision Loss For each identified segment, we calculate a unified attention map, $\boldsymbol{A}_i$, averaging attention across layers and relevant N text tokens:

$$\boldsymbol{A}_i = \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{N} [\![T_i \in \mathcal{T}_j]\!] \mathbf{CA}_i^l, \tag{3}$$

where $[\![\cdot]\!]$ is the Iverson bracket, $\mathbf{CA}_i^l$ is the cross-attention map for token i in layer l, and $\mathcal{T}_j$ denotes the set of tokens associated with the j-th segment.

The model is trained to predict noise for image-text pairs concatenated based on the parsed and segmented results. An additional loss term, designed to ensure focused reconstruction in areas relevant to each text-derived concept, is introduced. This loss is calculated as the Mean Squared Error (MSE) deviation from predefined masks corresponding to the segmented regions:

$$\mathcal{L}_{\text{ca}} = \mathbb{E}_{z,t}\left[\|\boldsymbol{A}_i(v_i, z_t) - M_i\|_2^2\right], \tag{4}$$

where $\boldsymbol{A}_i(v_i, z_t)$ is the cross-attention map between token v_i and noisy latent z_t, and M_i represents the mask for the i-th segment, which is derived from the segmented regions in our augmented dataset and appropriately resized to match the dimensions of the cross-attention maps.

Masked Diffusion Loss To ensure fidelity to the specified conditions, we apply a condition-selective diffusion loss that concentrates the denoising effort on conceptually significant regions. This focused loss function is applied solely to pixels within the regions delineated by the concept masks, which are derived from the non-zero features of the input structural conditions. Specifically, the masks are binary where non-zero feature areas are assigned a value of one, and areas lacking features are set to zero. Because of the sparsity of pose features for this condition, we use the all-ones mask. These masks serve to underscore the regions referenced in the corresponding text prompts:

$$\mathcal{L}_{\text{mask}} = \mathbb{E}_{z,\epsilon,t}\left[\|(\epsilon - \epsilon_\theta(z_t, t)) \odot M\|_2^2\right], \tag{5}$$

where M represents the union of binary mask obtained from input conditions, z_t denotes the noisy latent at timestep t, ϵ the injected noise, and ϵ_θ the estimated noise from the denoising network (U-Net).

The total loss function employed is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ldm}} + \lambda_{\text{ca}}\mathcal{L}_{\text{ca}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}, \tag{6}$$

with λ_{rec} and λ_{attn} set to 0.01. The integration of $\mathcal{L}_{\text{ca}}$ and $\mathcal{L}_{\text{mask}}$ ensure the model will focus at reconstructing the conditional region and attend to guided regions during generation.

3 Experiments

3.1 Datasets

In pursuit of our objective of achieving controlled Text-to-Image (T2I) generation, we employed the LAION improved_aesthetics_6plus (Schuhmann et al., 2022) dataset for our model training. Specifically, we meticulously curated a subset comprising 5,082,236 instances, undertaking the elimination of duplicates and applying filters based on criteria such as resolution and NSFW score. Given the targeted nature of our controlled generation tasks, the assembly of training data involved considerations of additional input conditions, specifically edge maps, sketch maps, depth maps, segmentation maps, and pose maps. The extraction of features from these maps adhered to the methodology expounded in(Zhang and Agrawala, 2023).

3.2 Experimental Setup

Structural Input Condition Extraction We start from the processing of various local conditions used in our experiments. To facilitate a comprehensive evaluation, we have incorporated a diverse range of structural conditions. These conditions include edge maps (Canny, 1986; Xie and Tu, 2015; Gu et al., 2022a), sketch maps (Simo-Serra et al., 2016), pose information (Cao et al., 2017), depth maps (Ranftl et al., 2020), and segmentation maps (Xiao et al., 2018), each extracted using specialized techniques. These conditions are crucial for guiding the text-to-image generation process, enabling FlexEControl to produce images that are both visually appealing and semantically aligned with the text prompts and structural inputs. The additional details for extracting those conditions are given in the Appendix.

Table 1: Text-to-image generation efficiency comparison: FlexEControl shows substantial reductions in memory cost, trainable parameters, and training time, highlighting its improved training efficiency with the same model architecture. Training times are averaged over three runs up to 400 iterations for consistency.

Models	Memory Cost ↓	# Params. ↓	Training Time ↓
Uni-ControlNet (Zhao et al., 2023)	20.47GB	1271M	5.69 ± 1.33s/it
LoRA (Hu et al., 2021)	17.84GB	1074M	3.97 ± 1.27 s/it
PHM (Zhang et al., 2021a)	15.08GB	819M	3.90 ± 2.01 s/it
FlexEControl (**ours**)	**14.33GB**	**750M**	**2.15 ± 1.42 s/it**

Evaluation Metrics We employ a comprehensive benchmark suite of metrics including mIoU (Rezatofighi et al., 2019), SSIM (Wang et al., 2004), mAP, MSE, FID (Heusel et al., 2017), and CLIP Score (Hessel et al., 2021; Radford et al., 2021) [1].

Baselines In our comparative evaluation, we assess T2I-Adapter (Mou et al., 2023), PHM (Zhang et al., 2021a), Uni-ControlNet (Zhao et al., 2023), and LoRA (Hu et al., 2021). The implementation details are given in the Appendix.

Implementation Details In accordance with the configuration employed in Uni-ControlNet, we utilized Stable Diffusion 1.5 [2] as the foundational model. Our model underwent training for a singular epoch, employing the AdamW optimizer (Kingma and Ba, 2014) with a learning rate set at 10^{-5}. Throughout all experimental iterations, we standardized the dimensions of input and conditional images to 512×512. The fine-tuning process was executed on P3 AWS EC2 instances equipped with 64 NVIDIA V100 GPUs.

For baseline implementations, we compare FlexEControl with T2I-Adapter (Mou et al., 2023), PHM (Zhang et al., 2021a), Uni-ControlNet (Zhao et al., 2023), and LoRA (Hu et al., 2021) where we implement LoRA and PHM layers over the trainable modules in Uni-ControlNet interms of generated image quality and controllability. The rank of LoRA is set to 4. For PHM (Zhang et al., 2021a), we implement it by performing Kronecker decomposition and share weights across different layer, with the number of decomposed matrix being 4.

For quantitative assessment, a subset comprising 10,000 high-quality images from the LAION improved_aesthetics_6.5plus dataset was utilized. The resizing of input conditions to 512×512 was conducted during the inference process.

The framework is further extended to accommodate video generation. The results can be found in the Appendix. In training the controllable video generation model with multiple input conditions, a straightforward strategy is employed to mask out conditions during the training process. In each iteration, a random sample, denoted as N_s, is drawn from $[1, N]$, where N is the number of total frames, to determine the number of frames that will incorporate the conditions. Subsequently, N_s unique values are drawn from the set and the conditions are retained for the corresponding frames.

3.3 Quantitative Results

Table 1 highlights FlexEControl's superior efficiency compared to Uni-ControlNet. It achieves a 30% reduction in memory cost, lowers trainable parameters by 41% (from 1271M to 750M), and significantly reduces training time per iteration from 5.69s to 2.15s.

Table 2 provides a comprehensive comparison of FlexEControl's performance against Uni-ControlNet and T2IAdapter across diverse input conditions. After training on a dataset of 5M text-image pairs, FlexEControl demonstrates better, if not superior, performance metrics compared to Uni-ControlNet and T2IAdapter. Note that Uni-ControlNet is trained on a much larger dataset (10M text-image pairs from the LAION dataset). Although there is a marginal decrease in SSIM scores for sketch maps and mAP scores for poses, FlexEControl

[1] https://github.com/jmhessel/clipscore
[2] https://huggingface.co/runwayml/stable-diffusion-v1-5

Table 2: Quantitative evaluation of controllability and image quality for single structural conditional inputs. FlexEControl performs overall better while maintaining much improved efficiency.

Models	Canny (SSIM)↑	MLSD (SSIM)↑	HED (SSIM)↑	Sketch (SSIM)↑	Depth (MSE)↓	Segmentation (mIoU)↑	Poses (mAP)↑	FID↓	CLIP Score↑
T2IAdapter (Mou et al., 2023)	0.4480	-	-	0.5241	90.01	0.6983	**0.3156**	27.80	0.4957
Uni-Control (Qin et al., 2023)	0.4977	0.6374	0.4885	0.5509	90.04	0.7143	0.2083	27.80	0.4899
Uni-ControlNet (Zhao et al., 2023)	0.4910	0.6083	0.4715	0.5901	90.17	0.7084	0.2125	27.74	0.4890
PHM (Zhang et al., 2021a)	0.4365	0.5712	0.4633	0.4878	91.38	0.5534	0.1664	27.91	0.4961
LoRA (Hu et al., 2021)	0.4497	0.6381	**0.5043**	0.5097	89.09	0.5480	0.1538	27.99	0.4832
FlexEControl (**ours**)	**0.4990**	**0.6385**	0.5041	0.5518	90.93	**0.7496**	0.2093	**27.55**	**0.4963**
ControlNet (Zhang and Agrawala, 2023)	0.4989	0.6172	0.4990	**0.6013**	**89.08**	0.7481	0.2024	27.62	0.4931

Table 3: Quantitative evaluation of controllability and image quality on FlexEControl along with its variants and Uni-ControlNet. For Uni-ControlNet, we implement multiple conditioning by adding two homogeneous conditional images after passing them through the feature extractor.

	Models	Canny (SSIM)↑	MLSD (SSIM)↑	HED (SSIM)↑	Sketch (SSIM)↑	Depth (MSE)↓	Segmentation (mIoU)↑	Poses (mAP)↑	FID↓	CLIP Score↑
Single Conditioning	Uni-ControlNet	0.3268	0.4097	0.3177	0.4096	98.80	0.4075	**0.1433**	29.43	0.4844
	FlexEControl (w/o L_{ca})	0.3698	0.4905	0.3870	0.4855	94.90	0.4449	0.1432	28.03	0.4874
	FlexEControl (w/o L_{mask})	0.3701	0.4894	0.3805	**0.4879**	**94.30**	0.4418	0.1432	28.19	0.4570
	FlexEControl	**0.3711**	**0.4920**	**0.3871**	0.4869	94.83	**0.4479**	0.1432	**28.03**	**0.4877**
Multiple Conditioning	Uni-ControlNet	0.3078	0.3962	0.3054	0.3871	98.84	0.3981	0.1393	28.75	0.4828
	FlexEControl (w/o L_{ca})	0.3642	0.4901	0.3704	0.4815	94.95	0.4368	0.1405	28.50	0.4870
	FlexEControl (w/o L_{mask})	0.3666	0.4834	0.3712	0.4831	94.89	0.4400	0.1406	28.68	0.4542
	FlexEControl	**0.3690**	**0.4915**	**0.3784**	**0.4849**	**92.90**	**0.4429**	**0.1411**	**28.24**	**0.4873**

excels in other metrics, notably surpassing Uni-ControlNet and T2IAdapter. This underscores our method's proficiency in enhancing efficiency and elevating overall quality and accuracy in controllable text-to-image generation tasks.

To validate FlexEControl's effectiveness in handling multiple structural conditions, we compared it with Uni-ControlNet through human evaluations. Two scenarios were considered: multiple homogeneous input conditions (300 images, each generated with 2 canny edge maps) and multiple heterogeneous input conditions (500 images, each generated with 2 randomly selected conditions). Results, summarized in Table 4, reveal that FlexEControl was preferred by 64.00% of annotators, significantly outperforming Uni-ControlNet (23.67%). This underscores FlexEControl's proficiency with complex, homogeneous inputs. Additionally, FlexEControl demonstrated superior alignment with input conditions (67.33%) compared to Uni-ControlNet (23.00%). In scenarios with random heterogeneous conditions, FlexEControl was preferred for overall quality and alignment over Uni-ControlNet.

In addition to our primary comparisons, we conducted an additional quantitative evaluation of FlexEControl and Uni-ControlNet. This evaluation focused on assessing image quality under scenarios involving multiple conditions from both the homogeneous and heterogeneous modalities. The findings of this evaluation are summarized in Table 5. FlexEControl consistently outperforms Uni-ControlNet in both categories, demonstrating lower FID scores for better image quality and higher CLIP scores for improved alignment with text prompts.

3.3.1 Ablation Studies

To substantiate the efficacy of FlexEControl in enhancing training efficiency while upholding commendable model performance, and to ensure a fair comparison, an ablation study was conducted by training models on an identical dataset. We trained FlexEControl along its variants and Uni-ControlNet on a subset of 100,000 training samples from LAION improved_aesthetics_6plus. When trained with the identical data, FlexEControl performs better than Uni-ControlNet. The outcomes are presented in Table 3. Evidently, FlexEControl exhibits substantial improvements over Uni-ControlNet when trained on the same dataset. This underscores the effectiveness of our approach in optimizing data utilization, concurrently diminishing computational costs, and enhancing efficiency in the text-to-image generation process.

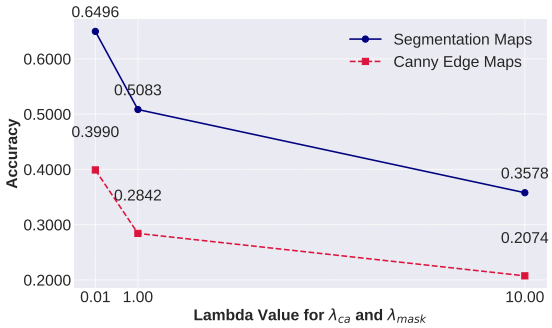

Figure 4: Ablation of lambda. FlexEControl performs the best when both $\lambda_{\text{ca}} = 0.01$ and $\lambda_{\text{mask}} = 0.01$.

Figure 5: Qualitative comparison of FlexEControl and existing controllable diffusion models with single condition. Text prompt: `A bed.` The image quality of FlexEControl is comparable to existing methods and Uni-ControlNet + LoRA, while FlexEControl has much more efficiency.

We also study the impact of λ_{ca} and λ_{mask} trained on the subset of 100,000 samples from LAION improved_aesthetics_6plus for 6,000 steps. We evaluated the score on SSIM of canny edge maps and mIoU of segmentation maps, results are shown in Figure 4. As observed, FlexEControl achieves optimal performance when both $\lambda_{\text{ca}} = 0.01$ and $\lambda_{\text{mask}} = 0.01$. Additionally, as indicated in Table **??**, FlexEControl outperforms the configurations where either λ_{ca} or λ_{mask} is set to zero, demonstrating the importance of incorporating both the cross-attention supervision loss and the masked diffusion loss. However, higher values of λ_{ca} and λ_{mask} could lead to instability during training, causing the model to focus excessively on local conditions, which results in worse performance.

3.3.2 Additional Results on Stable Diffusion 2

In our efforts to explore the versatility and adaptability of FlexEControl, we conducted additional experiments using the Stable Diffusion 2.1 model, available at Hugging Face's Model Hub. The results from these experiments are depicted in Table 6. FlexEControl can leverage the advancements in Stable Diffusion 2.1 to achieve even better performance in text-to-image generation tasks. For the sake of a fair comparison in the main paper, we conduct experiments using Stable Diffusion 1.5 model.

Figure 6: Qualitative comparison of FlexEControl and existing controllable diffusion models with multiple heterogeneous conditions. First row: FlexEControl effectively integrates both the segmentation and edge maps to generate a coherent image while Uni-ControlNet and LoRA miss the segmentation map and Uni-Control generates a messy image. Second row: The input condition types are one depth map and one sketch map. FlexEControl can do more faithful generation while all three others generate the candle in the coffee.

Table 4: Human evaluation of FlexEControl and Uni-ControlNet under homogenous and heterogeneous structural conditions, assessing both human preference and condition alignment. "Win" indicates FlexEControl's preference, "Tie" denotes equivalence, and "Lose" indicates Uni-ControlNet's preference. Results indicate that under homogeneous conditions, FlexEControl outperforms Uni-ControlNet in both human preference and condition alignment.

Condition Type	Metric	Win	Tie	Lose
Homogeneous	Human Preference (%)	**64.00**	12.33	23.67
	Condition Alignment (%)	**67.33**	9.67	23.00
Heterogeneous	Human Preference (%)	**9.80**	87.40	2.80
	Condition Alignment (%)	**6.60**	89.49	4.00

Table 5: Quantitative evaluation of controllability and image quality in scenarios with multiple conditions from heterogeneous and homogeneous modalities for FlexEControl and Uni-ControlNet. The 'heterogeneous' category averages the performance across one Canny condition combined with six other different modalities. The 'homogeneous' category represents the average performance across seven identical modalities (three inputs).

Condition Type	Baseline	FID↓	CLIP Score↑
Heterogeneous	Uni-ControlNet	27.81	0.4869
	FlexEControl	**27.47**	**0.4981**
Homogeneous	Uni-ControlNet	28.98	0.4858
	FlexEControl	**27.65**	**0.4932**

3.4 Qualitative Results

We present qualitative results of our FlexEControl under three different settings: single input condition, multiple heterogeneous conditions, and multiple homogeneous conditions, illustrated in Figure 5, Figure 6,

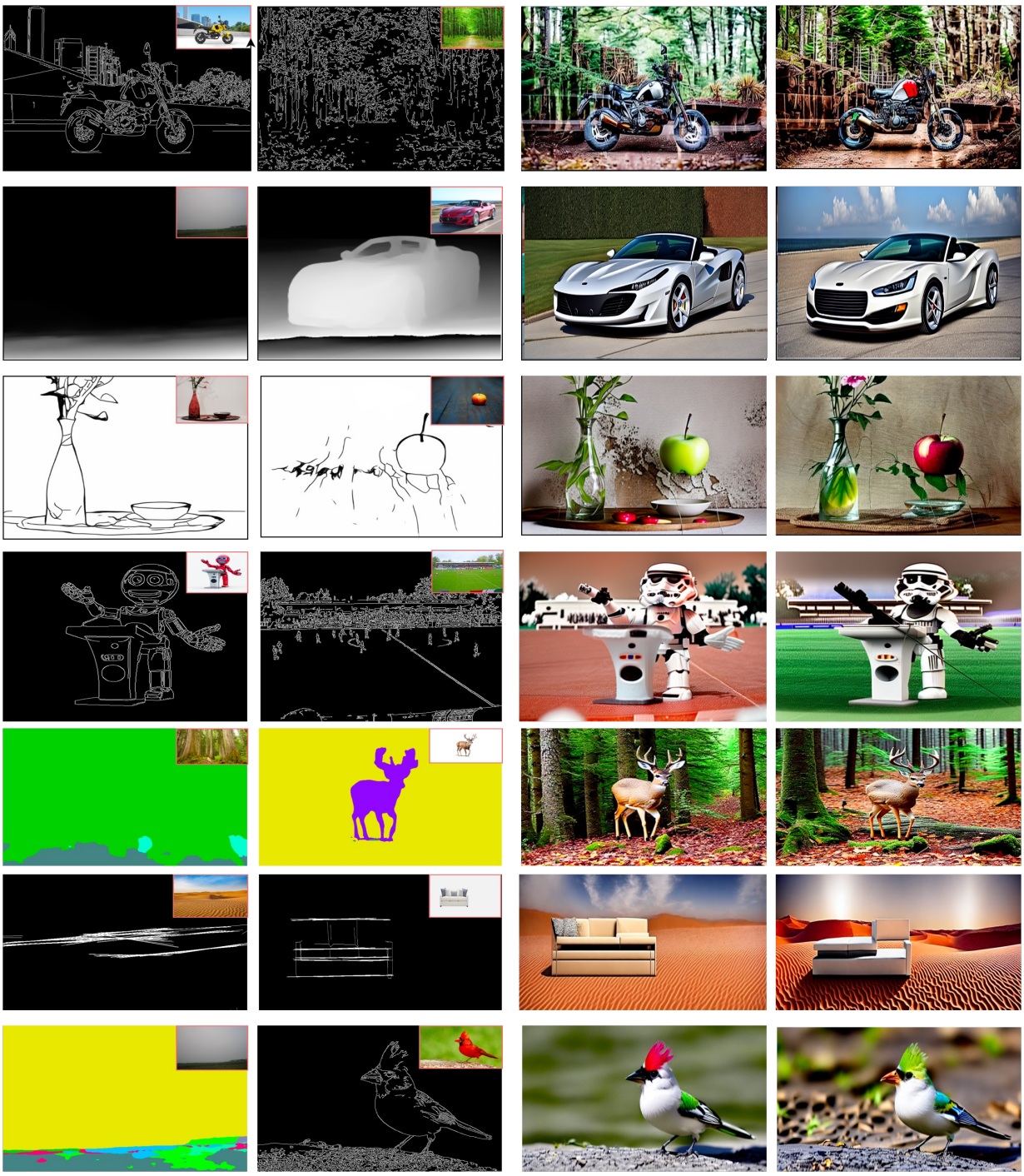

Figure 7: Qualitative performance of FlexEControl when conditioning on diverse compositions of multiple modalities. Each row in the figure corresponds to a unique type of condition, with the text prompts and conditions as follows: (first row) two canny edge maps with the prompt `A motorcycle in the forest`, (second row) two depth maps for `A car`, (third row) two sketch maps depicting `A vase with a green apple`, (fourth row) two canny edge maps for `Stormtrooper's lecture at the football field`, (fifth row) two segmentation maps visualizing `A deer in the forests`, (sixth row) two MLSD edge maps for `A sofa in a desert`, and (seventh row) one segmentation map and one edge map for `A bird`. These examples illustrate the robust capability of FlexEControl to effectively utilize multiple multimodal conditions, generating images that are not only visually compelling but also faithfully aligned with the given textual descriptions and input conditions.

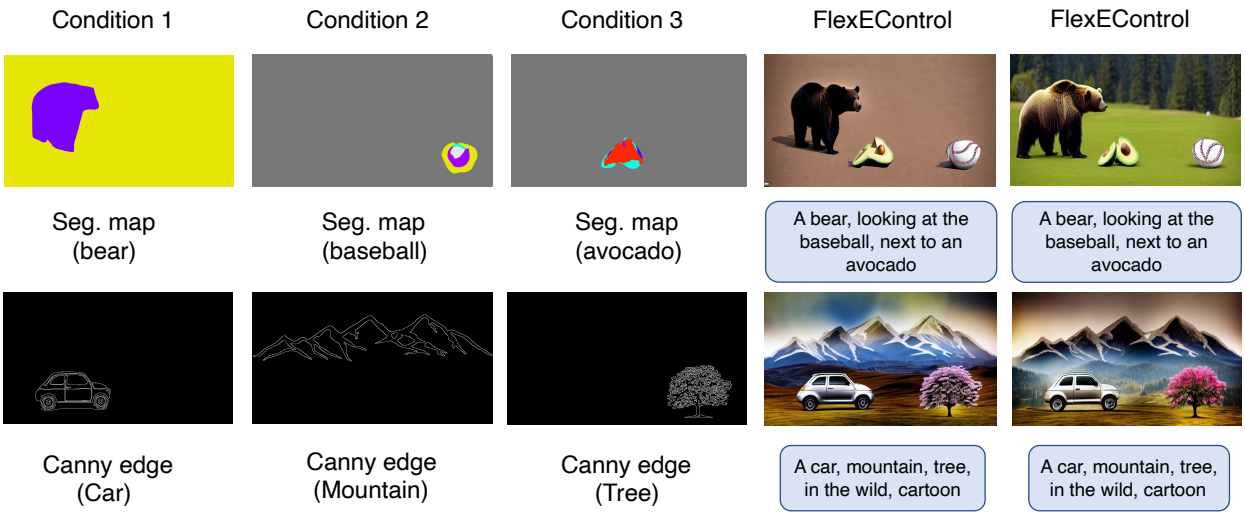

Figure 8: Qualitative results on multimodal control using three homogeneous condition images

Table 6: Quantitative evaluation of controllability and image quality trained on a subset of 100,000 samples. Human poses are evaluated solely within portrait images.

Models	Canny (SSIM)↑	MLSD (SSIM)↑	HED (SSIM)↑	Sketch (SSIM)↑	Depth (MSE)↓	Segmentation (mIoU)↑	Poses (mAP)↑	FID↓	CLIP Score↑
FlexEControl	0.3711	0.4920	0.3871	0.4869	94.83	0.4479	0.1432	28.03	0.4877
FlexEControl-SD 2.1	**0.3891**	**0.5273**	**0.4077**	**0.4960**	**93.58**	**0.4490**	**0.1562**	**25.08**	**0.5833**

and Figure 7, respectively. The results indicate that FlexEControl is comparable to baseline models when a single condition is input. However, with multiple conditions, FlexEControl consistently and noticeably outperforms other models. Particularly, under multiple homogeneous conditions, FlexEControl excels in generating overall higher quality images that align more closely with the input conditions, surpassing other models.

Furthermore, we show in Figure 8 the qualitative results for FlexEControl on mulimodal control using three homogeneous conditiona images. Further highlighting the versatility and effectiveness of FlexEControl in handling multiple conditions. Additionally, we showcase the extensibility of FlexEControl in controllable video generation. The results are presented in Figure 10 and Figure 11 in the Appendix, where results for providing one condition and multiple conditions are demonstrated.

3.5 Limitations

Although FlexEControl achieves strong performance, it shares the inherent limitations of diffusion-based image generation models. Notably, its effectiveness is constrained by the quality and diversity of the training data, which is drawn from a subset of the LAION dataset. This dataset exhibits certain biases, which can lead to suboptimal performance in specific scenarios. Some failure cases where FlexEControl struggled are illustrated in Figure **??**. FlexEControl could benefit from a more robust pretrained text-to-image generation backbone and could be further improved with access to better open-source datasets that mitigate the risk of generating biased, toxic, sexualized, or other harmful content.

4 Related Work

FlexEControl is an instance of efficient training and controllable text-to-image generation. Here, we overview modeling efforts in the subset of efficient training towards reducing parameters and memory cost and controllable T2I.

Efficient Training Prior work has proposed efficient training methodologies both for pretraining and fine-tuning. These methods have established their efficacy across an array of language and vision tasks. One of these explored strategies is Prompt Tuning (Lester et al., 2021), where trainable prompt tokens are appended to pretrained models (Schick and Schütze, 2020; Ju et al., 2021; Jia et al., 2022). These tokens can be added exclusively to input embeddings or to all intermediate layers (Li and Liang, 2021), allowing for nuanced model control and performance optimization. Low-Rank Adaptation (LoRA) (Hu et al., 2021) is another innovative approach that introduces trainable rank decomposition matrices for the parameters of each layer. LoRA has exhibited promising fine-tuning ability on large generative models, indicating its potential for broader application. Furthermore, the use of Adapters inserts lightweight adaptation modules into each layer of a pretrained transformer (Houlsby et al., 2019; Rücklé et al., 2021). This method has been successfully extended across various setups (Zhang et al., 2021b; Gao et al., 2021; Mou et al., 2023), demonstrating its adaptability and practicality. Other approaches including post-training model compression (Fang et al., 2023) facilitate the transition from a fully optimized model to a compressed version – either sparse (Frantar and Alistarh, 2023), quantized (Li et al., 2023a; Gu et al., 2022b), or both. This methodology was particularly helpful for parameter quantization (Dettmers et al., 2023). Different from these methodologies, FlexEControl puts forth a new unified strategy that aims to enhance the efficient training of text-to-image diffusion models through the leverage of low-rank structure. Yeh et al. (2023) and Marjit et al. (2024) explore the use of Kronecker Product for tuning T2I models. These approaches focus on enhancing control and efficiency, particularly in subject-driven or style-preserving generation tasks. FlexEControl is the first to leverage it for multimodal conditioning, leading to enhanced flexibility and efficiency in handling diverse input modalities. FlexEControl is also a general approach that can be applied to UniControl Qin et al. (2023) or other backbones.

Controllable Text-to-Image Generation Recent developments in the text-to-image generation domain strives for more control over image generation, enabling more targeted, stable, and accurate visual outputs, several models like T2I-Adapter (Mou et al., 2023) and Composer (Huang et al., 2023) have emerged to enhance image generations following the semantic guidance of text prompts and multiple different structural conditional control. However, existing methods are struggling at dealing with multiple conditions from the same modalities, especially when they have conflicts, e.g. multiple segmentation maps and at the same time follow the guidance of text prompts; Recent studies also highlight challenges in controllable text-to-image generation (T2I), such as omission of objects in text prompts and mismatched attributes (Lee et al., 2023; Bakr et al., 2023), showing that current models are strugging at handling controls from different conditions. Towards these, the Attend-and-Excite method Chefer et al. (2023) refines attention regions to ensure distinct attention across separate image regions. ReCo Yang et al. (2023), GLIGEN Li et al. (2023b), and Layout-Guidance Chen et al. (2023) allow for image generation informed by bounding boxes and regional descriptions. Mo et al. (2024) offers a training-free approach to multimodal control. FlexEControl improves the model's controllability by proposing a new training strategy, distinguishing itself by targeting the flexibility and efficiency of multimodal control, especially in scenarios with conflicting conditions from the same or different modalities (e.g., multiple segmentation maps combined with text prompts).

5 Conclusion

In this work, we present FlexEControl, an approach designed to enhance both the flexibility and efficiency of controllable diffusion-based text-to-image generation. Our method introduces several key innovations: Dataset Augmentation with Text Parsing and Segmentation, Cross-Attention Supervision, and Masked Diffusion Loss, which together significantly improve the model's ability to handle diverse and conflicting multimodal inputs. Additionally, we propose an Efficient Training strategy that optimizes parameter, data, and memory efficiency without sacrificing performance or inference speed. We also demonstrate that sharing a common set of decomposed weights across different multimodal conditions via Kronecker Decomposition can further optimize parameter space and enhance efficiency. These findings suggest that FlexEControl can be readily adapted to other architectures, offering a scalable solution for future developments in text-to-image generation. Future work could explore more advanced decomposition techniques and their application to cutting-edge diffusion backbones or Diffusion Transformers (DiTs), aiming to further optimize model efficiency, complexity, and expressive power.

Broader Impact Statement

While FlexEControl demonstrates promising results in efficient and controllable text-to-image generation, the ability of FlexEControl to generate realistic images based on textual descriptions raises ethical concerns, especially regarding the creation of misleading or deceptive content. It is imperative to establish guidelines and ethical standards for the use of such technology to prevent misuse in generating deepfakes or propagating false information.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. *arXiv:2012.13255 [cs]*.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20041–20053.

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

Samy Bengio, Jason Weston, and David Grangier. 2010. Label embedding trees for large multi-class tasks. In *NIPS*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698.

Shidong Cao, Wenhao Chai, Shengyu Hao, Yanting Zhang, Hangyue Chen, and Gaoang Wang. 2023. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *IEEE Transactions on Multimedia*.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Minghao Chen, Iro Laina, and Andrea Vedaldi. 2023. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2023. Structural pruning for diffusion models. *arXiv preprint arXiv:2305.10924*.

Elias Frantar and Dan Alistarh. 2023. Massive language models can be accurately pruned in one-shot. *arXiv preprint arXiv:2301.00774*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

Geonmo Gu, Byungsoo Ko, SeoungHyun Go, Sung-Hyun Lee, Jingeun Lee, and Minchul Shin. 2022a. Towards light-weight and real-time line segment detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 726–734.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022b. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. 2020. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11020–11029.

Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. 2022. Parameter-efficient fine-tuning for vision transformers. *arXiv preprint arXiv:2203.16329*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. *arXiv:1902.00751 [cs, stat]*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685 [cs]*.

Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling Causal Effect of Data in Class-Incremental Learning. page 10.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2021. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, and Daniel Haziza. 2022. xformers: A modular and hackable transformer modelling library. `https://github.com/facebookresearch/xformers`.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. *arXiv:2101.00190 [cs]*.

Xiuyu Li, Long Lian, Yijiang Liu, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. 2023a. Q-diffusion: Quantizing diffusion models. *arXiv preprint arXiv:2302.04304*.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient Low-Rank Hypercomplex Adapter Layers. *arXiv:2106.04647 [cs]*.

Shyam Marjit, Harshit Singh, Nityanand Mathur, Sayak Paul, Chia-Mu Yu, and Pin-Yu Chen. 2024. Diffusekrona: A parameter efficient fine-tuning method for personalized diffusion model. *arXiv preprint arXiv:2402.17412*.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. 2024. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475.

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

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. 2018. Efficient neural architecture search via parameter sharing. In *ICML*.

Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. 2023. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021b. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the Efficiency of Adapters in Transformers. *arXiv:2010.11918 [cs]*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022a. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv:2205.11487*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *EMC2 @ NeurIPS*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

Edgar Simo-Serra, Satoshi Iizuka, Kazuma Sasaki, and Hiroshi Ishikawa. 2016. Learning to simplify: fully convolutional networks for rough sketch cleanup. *ACM Transactions on Graphics (TOG)*, 35(4):1–11.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP Architecture for Vision. *arXiv:2105.01601 [cs]*.

Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. 2022. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*.

Yixin Wang and Michael I. Jordan. 2021. Desiderata for Representation Learning: A Causal Perspective. *arXiv:2109.03795 [cs, stat]*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Yeming Wen, Dustin Tran, and Jimmy Ba. 2020. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434.

Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403.

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. 2023. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255.

Shih-Ying Yeh, Yu-Guan Hsieh, Zhidong Gao, Bernard BW Yang, Giyeong Oh, and Yanmin Gong. 2023. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In *The Twelfth International Conference on Learning Representations*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. *arXiv:2106.10199 [cs]*.

Aston Zhang, Yi Tay, Shuai Zhang, Alvin Chan, Anh Tuan Luu, Siu Cheung Hui, and Jie Fu. 2021a. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters. *arXiv preprint arXiv:2102.08597*.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021b. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. *arXiv:2111.03930 [cs]*.

Zhongping Zhang, Jian Zheng, Jacob Zhiyuan Fang, and Bryan A Plummer. 2023. Text-to-image editing by image information removal. *arXiv preprint arXiv:2305.17489*.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*.

This Appendix is organized as follows:

- Appendix A contains our detailed model architectures;
- Appendix B contains additional implementation details;
- Appendix C contains additional results;
- Appendix D contains additional study on text-to-image pretraining;
- Appendix E contains additional related works;
- Appendix F contains details for the human evaluation setup;

A Detailed Model Architecture

In ControlNet (Zhang and Agrawala, 2023) and Uni-ControlNet (Zhao et al., 2023), the weights of Stable Diffusion (SD) (Rombach et al., 2022) are fixed and the input conditions are fed into zero-convolutions and added back into the main Stable Diffusion backbone. Specifically, for Uni-ControlNet, they uses a multi-scale condition injection strategy that extracts features at different resolutions and uses them for condition injection referring to the implementation of Feature Denormalization (FDN):

$$\begin{aligned} &\mathrm{FDN}\left(Z, c\right) = \mathrm{norm}\left(Z\right) \cdot \left(1 + \Phi\left(\mathrm{zero}\left(h_r\left(c\right)\right)\right)\right) \\ &+ \Phi\left(\mathrm{zero}\left(h_r\left(c\right)\right)\right), \end{aligned} \tag{7}$$

where Z denotes noise features, c denotes the input conditional features, Φ denotes learnable convolutional layers, and zero denotes zero convolutional layer. The zero convolutional layer contains weights initialized to zero. This ensures that during the initial stages of training, the model relies more on the knowledge from the backbone part, gradually adjusting these weights as training progresses. The use of such layers aids in preserving the architecture's original behavior while introducing structure-conditioned inputs. We use the similar model architecture while we perform efficient training proposed in the main paper. We show the model architecture in Figure 9.

B Additional Implementation Details

In this section, we provide further details about the implementation aspects of our approach.

B.1 Additional Details of Structural Input Conditions Extraction

- **Edge Maps**: For generating edge maps, we utilized two distinct techniques:
  - Canny Edge Detector (Canny, 1986) - A widely used method for edge detection in images.
  - HED Boundary Extractor (Xie and Tu, 2015) - Holistically-Nested Edge Detection, an advanced technique for identifying object boundaries.
  - MLSD (Gu et al., 2022a) - A method particularly designed for detecting multi-scale line segments in images.
- **Sketch Maps**: We adopted a sketch extraction technique detailed in Simo-Serra et al. (2016) to convert images into their sketch representations.
- **Pose Information**: OpenPose (Cao et al., 2017) was employed to extract human pose information from images, which provides detailed body joint and keypoint information.
- **Depth Maps**: For depth estimation, we integrated Midas (Ranftl et al., 2020), a robust method for predicting depth information from single images.

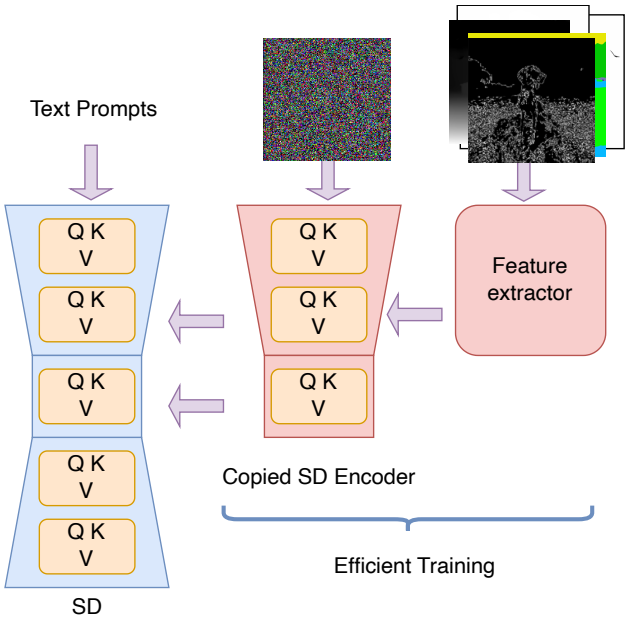

Figure 9: Detailed model architecture in FlexEControl. The Stable Diffusion part is fixed and others are trainable.

- **Segmentation Maps**: Segmentation of images was performed using the method outlined in Xiao et al. (2018), which focuses on accurately segmenting various objects within an image.

Each of these conditions plays a crucial role in guiding the text-to-image generation process, helping FlexEControl to generate images that are not only visually appealing but also semantically aligned with the given text prompts and structural conditions.

B.2 Additional Details of Evaluation Metrics

mIoU (Rezatofighi et al., 2019): Mean Intersection over Union, a metric that quantifies the degree of overlap between predicted and actual segmentation maps.

SSIM (Wang et al., 2004): Structural Similarity, a metric evaluating the structural similarity in generated outputs, applied to Canny edges, HED edges, MLSD edges, and sketches.

mAP: Mean Average Precision, utilized for pose maps, measuring the precision of localization across multiple instances.

MSE: Mean Squared Error, employed for depth maps, MSE quantifies the pixel-wise variance, providing an assessment of image fidelity.

FID (Heusel et al., 2017): Fréchet Inception Distance, which serves as a metric to quantify the realism and diversity of the generated images. A lower FID value indicates higher quality and diversity of the output images.

CLIP Score (Hessel et al., 2021; Radford et al., 2021): Employing CLIP Score, we gauge the semantic similarity between the generated images and the input text prompts.

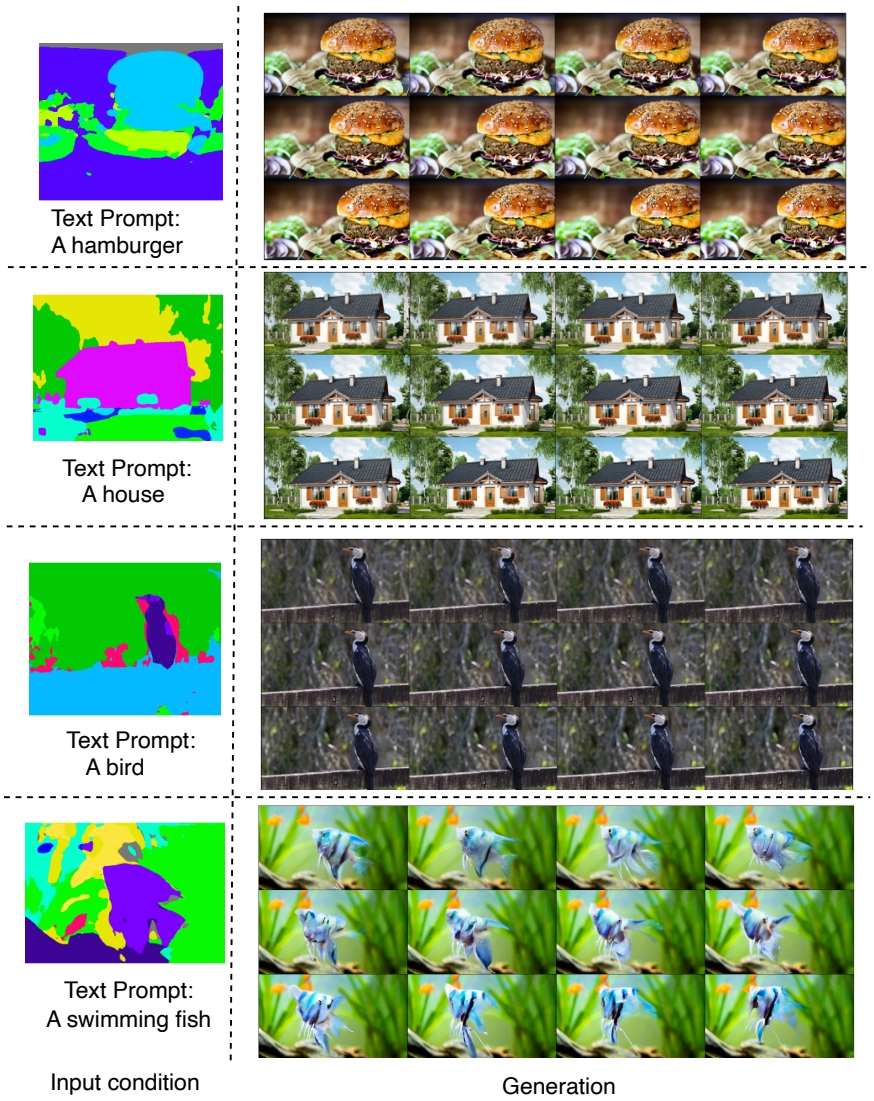

Figure 10: Results from FlexEControl on controllable text-to video generation (single condition).

C Additional Results

C.1 Additional Qualitative Results on Video Generation

FlexEControl can be further extended to accommodate video generation. In training the controllable video generation model with multiple input conditions, a straightforward strategy is employed to mask out conditions during the training process. In each iteration, a random sample, denoted as N_s, is drawn from $[1, N]$ to determine the number of frames that will incorporate the conditions. Subsequently, N_s unique values are drawn from the set $1, 2, ..., N$, and the conditions are retained for the corresponding frames.

In this section, we showcase the extensibility of FlexEControl in controllable video generation. The results are presented in Figure 10 and Figure 11, where results for providing one condition and multiple conditions are demonstrated.

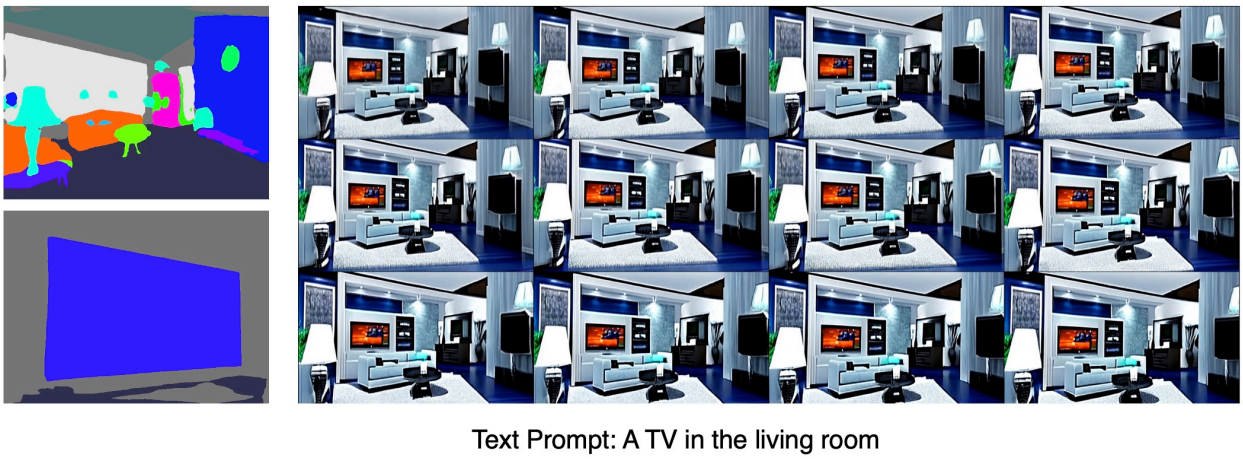

Figure 11: FlexEControl on using multiple conditions for video generation.

D Training a Small U-Net Backbone

In this section, we discuss further methods to refine the training of a lightweight Stable Diffusion backbone within FlexEControl, aiming to further curtail the number of trainable parameters and minimize memory usage end-to-end. The resulting pre-trained Stable Diffusion backbone, which we denote as FlexEControl-pretraining, offers a more lightweight alternative to the original model while retaining versatility for application in a variety of tasks.

Building upon the strategies delineated in the main paper, we architect a streamlined U-Net structure utilizing low-rank decomposition. This design is complemented by the implementation of knowledge distillation techniques throughout the training process to cultivate an efficient text-to-image generative model. Our training regimen unfolds in two distinct phases: Initially, we focus on establishing a lightweight T2I diffusion model founded on a conventional U-Net framework, with knowledge distillation enhancing this foundational stage. Subsequently, we move to fine-tuning introduced in the main paper, enabling the model to adeptly manage controlled T2I generation tasks. This bifurcated approach yields significant resource savings both in fine-tuning and in the overall model parameter count, setting a new benchmark for efficiency in generative modeling.

D.1 Background on Low-rank Training

Background on Training in Low-dimensional Space Let $\theta^D = \left[{\theta_0}^D \dots {\theta_m}^D\right]$ be a set of m D-dimensional parameters that parameterize the U-Net within the Stable Diffusion. Instead of optimizing the noise prediction loss in the original parameter space $\left(\theta^D\right)$, we are motivated to train the model in the lower-dimensional space $\left(\theta^d\right)$ (Aghajanyan et al., 2020). Our overall pipeline is trying to train the controllable text-to-image diffusion model in such a lower-dimension space to improve the overall efficiency.

An overview of our proposed two-stage pipeline is shown in Figure 14. We first train the U-Net of a text-to-image model with a low-rank schema. Specifically, we employ matrix factorization techniques that decompose high-dimensional matrices into smaller matrices, capturing essential features with reduced computational overhead. This process is augmented through knowledge distillation, visually represented in green on Figure 14. We then conduct efficient fine-tuning using the methods (shown in the yellow part on Figure 14) with the methods introduced in the main paper, where we employ low-rank decomposition and Kronecker decomposition to streamline the parameter space.

Low-rank Text-to-image Diffusion Model To establish a foundational understanding of our model, it's crucial to first comprehend the role of U-Nets in the diffusion process. In diffusion models, there exists

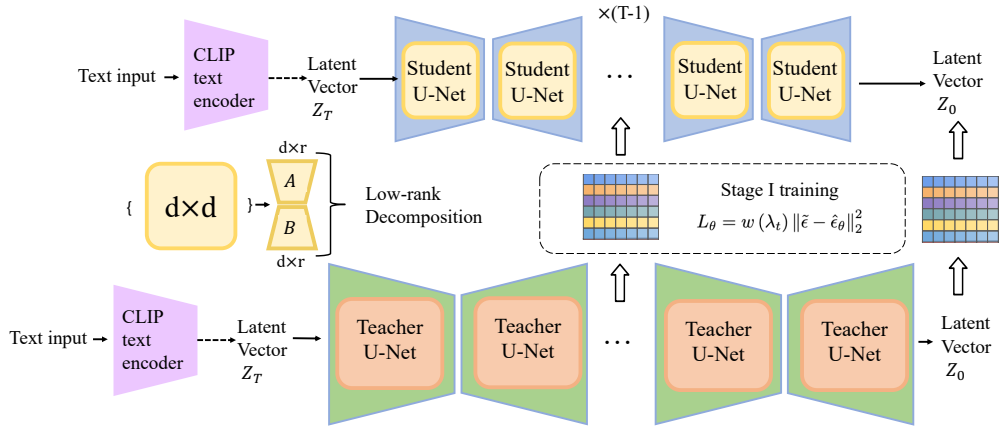

Figure 12: Overview of the Stage-1 training: Training a low-rank U-Net using knowledge distillation from a teacher model (green) to the student model (blue). This process involves initializing the student U-Net with a decomposition into low-rank matrices and minimizing the loss between the predicted noise representations from the student and teacher.

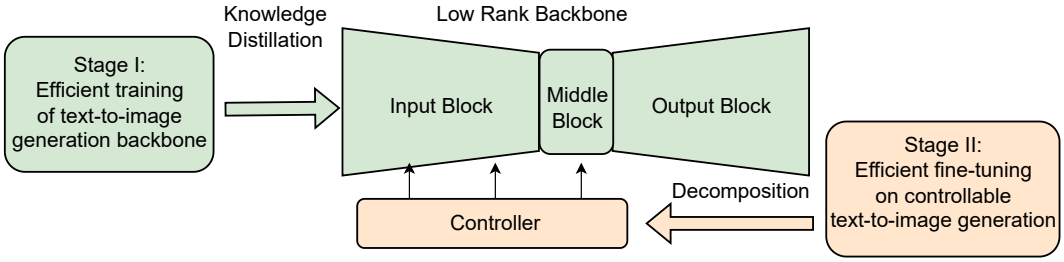

Figure 13: The overview pipeline of our method. Our method improves the efficiency of controllable text-to-image generation from two aspects. At pretraining stage, we propose an efficient pretraining method for the standard text-to-image generation via knowledge distillation. For the finetuning stage introduced in the main paper, we propose to resort to low-rank and Kronecker decomposition to reduce the tunable parameter space.

an input language prompt y that is processed by a encoder τ_θ. This encoder projects y to an intermediate representation $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$, where M is denotes the token length, and d_τ denotes the dimension of the embedding space . This representation is subsequently mapped to the intermediate layers of the U-Net through a cross-attention layer given by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V, \tag{8}$$

with $Q = \mathbf{W}_Q \varphi_i(z_t)$, $K = \mathbf{W}_K \tau_\theta(y)$, $V = \mathbf{W}_V \tau_\theta(y)$. In this context, $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon}$ is an intermediate representation of the U-Net. The terms $\mathbf{W}_V \in \mathbb{R}^{d \times d_\epsilon}, \mathbf{W}_Q \in \mathbb{R}^{d \times d_\tau}, \mathbf{W}_K \in \mathbb{R}^{d \times d_\tau}$ represent learnable projection matrices.

Shifting focus to the diffusion process, during the t-timestep, we can represent:

$$K = \mathbf{W}_K \tau_\theta(y) = AB\tau_\theta(y), \tag{9}$$
$$V = \mathbf{W}_V \tau_\theta(y) = AB\tau_\theta(y), \tag{10}$$

where A and B are decomposed low-rank matrices from the cross-attnetion matrices, d_τ and d_ϵ denote the dimension for the text encoder and noise space respectively. Conventionally, the diffusion model is trained via minimizing $\mathcal{L}_\theta = \left\|\epsilon - \epsilon_\theta\right\|_2^2$, where ϵ is the groundtruth noise and ϵ_θ is the predicted noise from the model.

Central to our strategy is a knowledge distillation process. This involves guiding a novice or 'Student' diffusion model using feature maps that draw upon the wisdom of a more seasoned 'Teacher' model. A pivotal insight

from our study lies in the mathematical congruence between the low-rank training processes across both training phases, unveiling the symmetries in low-rank training trajectories across both phases.

To fully exploit the prior knowledge from the pretrained teacher model while exploiting less data and training a lightweight diffusion model, we propose a new two-stage training schema. The first one is the initialization strategy to inherit the knowledge from the teacher model. Another is the knowledge distillation strategy. The overall pipeline is shown in Figure 12.

D.2 Initialization

Directly initializing the student U-Net is not feasible due to the inconsistent matrix dimension across the Student and teacher U-Net. We circumvent this by decomposing U-Net into two low-rank matrices, namely A and B for the reconstruction. We adopt an additional transformation to adapt the teacher's U-Net weights to the Student, which leverages the Singular Value Decomposition (SVD) built upon the teacher U-Net. The initialization process can be expressed as:

1. Compute the SVD of the teacher U-Net: Starting with the teacher U-Net parameterized by θ_0, we compute its SVD as $\theta_0 = U\Sigma V^T$.
2. Extract Low-Rank Components: to achieve a low-rank approximation, we extract the first k columns of U, the first k rows and columns of Σ, and the first k rows of V^T. This results in matrices U_k, Σ_k, and V_k^T as follows:

$$U_k = \text{first } k \text{ columns of } U, \tag{11}$$

$$\Sigma_k = \text{first } k \text{ rows \& columns of } \Sigma, \tag{12}$$

$$V_k^T = \text{first } k \text{ rows of } V^T \tag{13}$$

3. We then initialize the student U-Net with $U_k\Sigma_k$ and V_k^T that encapsulate essential information from the teacher U-Net but in a lower-rank format.

We observe in practice that such initialization effectively retains the prior knowledge inherited from Teacher U-Net while enabling the student U-Net to be represented in a compact form thus computationally more efficient for later training.

D.3 Loss Function

We propose to train our Student U-Net with knowledge distillation (Meng et al., 2023) to mimic the behavior of a teacher U-Net. This involves minimizing the loss between the student's predicted noise representations and those of the teacher. To be specific, our training objective can be expressed as:

$$\mathcal{L}_\theta = w\left(\lambda_t\right)\left\|\tilde{\epsilon} - \hat{\epsilon}_\theta\right\|_2^2, \tag{14}$$

where $\tilde{\epsilon}$ denotes the predicted noise in the latent space of Stable Diffusion from the teacher model, $\hat{\epsilon}_\theta$ is the corresponding predicted noise from the student model, parameterized by θ, and $w\left(\lambda_t\right)$ is a weighting function that may vary with the time step t. Such an objective encourages the model to minimize the squared Euclidean distance between the teacher and Student's predictions thus providing informative guidance to the Student. We also tried combining the loss with the text-to-image Diffusion loss but using our training objective works better.

D.4 Experimental Settings

In the pretraining stage, we used the standard training scheme of Stable Diffusion (Rombach et al., 2022) with the classifier-free guidance (Ho and Salimans, 2022). We employed the Stable Diffusion 2.1 [3] model in conjunction with xFormers (Lefaudeux et al., 2022) and FlashAttention (Dao et al., 2022) using the implementation available in HuggingFace Diffusers [4].

[3] https://huggingface.co/stabilityai/stable-diffusion-2-1
[4] https://huggingface.co/docs/diffusers/index

Table 7: Comparing U-Net models: Original, decomposed, with and without Knowledge Distillation. FlexEControl-Pretraining showcases a promising balance between performance and efficiency. Note that compared with Stable Diffusion, FlexEControl-Pretraining is only trained on 5 million data. FlexEControl-Pretraining beats Decomposed U-Net w/o Distillation interms of FID and CLIP Score, suggesting the effectiveness of our distillation strategy in training the decomposed U-Net.

Methods	FID↓	CLIP Score↑	# Parameters ↓
Stable Diffusion	27.7	0.824	1290M
Standard U-Net w/o Distill.	66.7	0.670	1290M
Decomposed U-Net w/o Distill.	84.3	0.610	790M
FlexEControl-Pretraining	45.0	0.768	790M

Table 8: Performance and resource metrics comparison of FlexEControl with the baseline Uni-ControlNet. The FlexEControl approach with distillation shows a significant reduction in resource consumption while providing competitive image quality and outperforming in controllability metrics, especially in segmentation maps. The Δ column shows the improvement of FlexEControl (w/o distillation) compared with no distillation.

	Metrics	Uni-ControlNet	FlexEControl		Δ
			w/o Distill.	w/ Distill.	
Efficiency	Memory Cost ↓	20GB	**11GB**	**11GB**	**0**
	# Params. ↓	1271M	**536M**	**536M**	**0**
Image Quality	FID ↓	27.7	84.0	43.7	**- 40.3**
	CLIP Score ↑	0.82	0.61	0.77	**+0.16**
Controllability	Sketch Maps (CLIP Score)↑	0.49	0.40	0.46	**+0.06**
	Edge Maps (NMSE) ↓	0.60	0.54	0.57	**+0.03**
	Segmentation Maps (IoU) ↑	0.70	0.40	0.74	**+0.34**

D.5 Results

Table 7 illustrates the comparison between different variations of our method in the pretraining stage, including original U-Net, decomposed low-rank U-Net, and their respective performance with and without knowledge distillation. It is observed that the decomposed low-rank U-Net models demonstrate efficiency gains, with a reduction in the total number of parameters to 790M, although at the cost of some fidelity in metrics such as FID and CLIP Score. Employing distillation helps to mitigate some of these performance reductions.

Table 8 illustrates the comparison between FlexEControl including pretraining and the baseline training end-to-end. It is observed that the decomposed low-rank U-Net models demonstrate efficiency gains, with a reduction in the total number of parameters to 536M, although at the cost of some fidelity in metrics such as FID and CLIP Score. Employing distillation helps to mitigate some of these performance reductions.

These collective results affirm our method's capability to not only enhance efficiency but also improve or maintain performance across various aspects of text-to-image generation.

E Additional Related Works

Knowledge Distillation for Vision-and-Language Models Knowledge distillation (Gou et al., 2021), as detailed in prior research, offers a promising approach for enhancing the performance of a more streamlined "student" model by transferring knowledge from a more complex "teacher" model (Hinton et al., 2015; Sanh et al., 2019; Hu et al.; Gu et al., 2021; Li et al., 2021). The crux of this methodology lies in aligning the

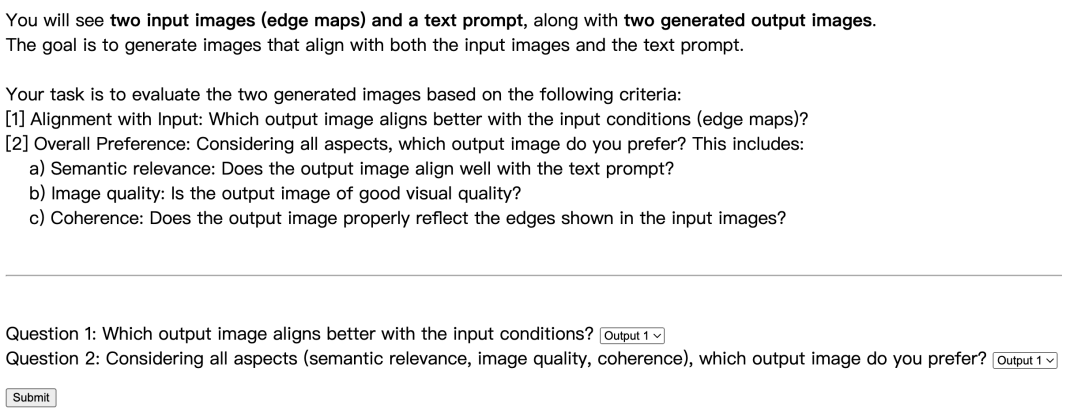

Figure 14: Screenshot for human evaluation tasks on the Amazon Mechanical Turk crowdsource evaluation platform.

predictions of the student model with those of the teacher model. While a significant portion of existing knowledge distillation techniques leans towards employing pretrained teacher models (Tolstikhin et al., 2021), there has been a growing interest in online distillation methodologies (Wang and Jordan, 2021). In online distillation (Guo et al., 2020), multiple models are trained simultaneously, with their ensemble serving as the teacher. Our approach is reminiscent of online self-distillation, where a temporal and resolution ensemble of the student model operates as the teacher. This concept finds parallels in other domains, having been examined in semi-supervised learning (Peters et al., 2017), label noise learning (Bengio et al., 2010), and quite recently in contrastive learning (Chen et al., 2020). Our work on distillation for pretrained text-to-image generative diffusion models distinguishes our method from these preceding works. (Salimans and Ho, 2022; Meng et al., 2023) propose distillation strategies for diffusion models but they aim at improving inference speed. Our work instead aims to distill the intricate knowledge of teacher models into the student counterparts, ensuring both the improvements over training efficiency and quality retention.

F Human Evaluation Interface

We give the human evaluation interface in Figure 14. The human evaluators are mainly asked to finish two tasks and choose their preference from three perspectives.