Controlled FastComposer

Alex Hu
Massachusetts Institute of Technology
alexhu@mit.edu

Sean Li
Massachusetts Institute of Technology
seanjli@mit.edu

Emily Liu
Massachusetts Institute of Technology
emizfliu@mit.edu

Sarah Zhang
Massachusetts Institute of Technology
sjzhang@mit.edu

Abstract

Recent advancements in generative models, specifically diffusion models, have demonstrated their ability to produce high-quality, diverse images. However, challenges persist, including inefficiencies in subject-specific fine-tuning, difficulties in multi-subject generation due to feature blending, and tendencies towards subject over-fitting. Although FastComposer has addressed some of these issues by incorporating subject embeddings, cross-attention localization, and delayed subject conditioning, it falls short in providing users with fine-grained spatial control over generated images, especially in terms of poses and expressions. To overcome this limitation, we present Controlled FastComposer, an integration of FastComposer with ControlNet. ControlNet extends diffusion models by introducing an additional layer of conditioning, which can involve edge detection or human pose detection. This offers users enhanced flexibility in image generation while maintaining the desired poses. Our approach demonstrates that Controlled FastComposer effectively addresses the control limitations inherent in FastComposer, pushing the boundaries of image generation control and efficiency. In particular, Controlled FastComposer surpasses the performance observed with ControlNet alone or with undelayed conditioning. The proposed integration has the potential to unlock new avenues for creative expression, providing users with the opportunity to create controlled visual content.

1 Introduction

Within the landscape of image generation, diffusion models have emerged as powerful tools [1, 2, 3]. This new class of generative models, driven by the training of denoising functions, excels at generating high-quality image samples from random noise. Despite their notable success, diffusion models encounter challenges related to subject-specific fine-tuning, identity blending, and subject overfitting. Specifically, the generated images struggle to accurately represent the distinctive features of each subject.

In response to these challenges, FastComposer emerges as a promising solution for multi-subject text-to-image generation, eliminating the need for extensive fine-tuning [7]. However, FastComposer is limited in the level of control afforded to the user over the creative process. For example, users face constraints in the spatial composition of the image, such as manipulating the poses or expressions of the input image subject within the generated output image [8].

To enhance user control over the creative process, we introduce Controlled FastComposer, which combines the capabilities of FastComposer with ControlNet to empower users with fine-grained spatial control over the generated images. By drawing upon the strengths of these two frameworks,

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Controlled FastComposer seeks to bridge the existing gap, opening up new possibilities for users to exert control over the spatial composition, poses, and expressions within the domain of multi-subject text-to-image generation.

2 Related Work

FastComposer. FastComposer introduces a novel approach that facilitates streamlined and personalized multi-subject text-to-image generation without the need for extensive fine-tuning. FastComposer replaces generic text tokens with subject embeddings derived from referenced images, which allows for image generation based on subject-augmented conditioning with only forward passes. Beyond reducing the computational overhead of the image generation process, FastComposer introduces two key techniques: cross-attention localization and delayed subject conditioning. First, cross-attention localization is proposed to address the challenges posed by identity blending, where the model blends distinct characteristics among different subjects. This method guides the model to map subject features to distinct regions of the image. Second, delayed subject conditioning is proposed to address the challenges posed by subject overfitting, where the model overfits to the input image and disregards textual instructions. This method incorporates text-only conditioning during the initial denoising stage to establish the image structure and subject-augmented conditioning in the remaining denoising steps to enhance the subject appearance. As a result, FastComposer sets a new standard for efficient and personalized content creation [7].

ControlNet. ControlNet is a neural network architecture designed to augment large, pretrained text-to-image diffusion models with spatial conditioning controls. The motivation behind this arises from the inherent limitations of existing text-to-image models, which struggle to afford users with precise control over the spatial composition, including layouts, poses, shapes, and forms. To address this challenge, ControlNet leverages the concept of conditioning inputs, such as edge maps, human pose skeletons, and segmentation maps, as valuable guides for the image generation process. Although the datasets for particular conditions, like object shape or human pose extraction, are considerably smaller than the datasets used for general text-to-image training, ControlNet locks the parameters of a large, pretrained diffusion model while creating a trainable copy of its encoding layers. This effectively treats the pretrained model as a robust backbone for learning conditional controls, with the trainable copy connected to the original model through "zero convolution layers", ensuring the model's stability and scalability across different datasets [8].

3 Methods

Model Architecture. Controlled FastComposer consists of two integral components: the trained FastComposer denoiser and the ControlNet guidance structure attached to each encoder block. A traditional use case for the ControlNet guidance architecture is using a Stable Diffusion U-Net, which gives the ability for the user to condition denoised outputs on spatial controls. On the other hand, FastComposer consists of a modified U-Net, trained with text tokens augmented with subject embeddings from reference images whose cross-attention maps are masked during training. Controlled FastComposer merges these two models, denoising with a pretrained FastComposer U-Net, which takes subjected-embedding textual inputs, and processing image latents prepared using a pretrained ControlNet model.

As shown in Figure 1, given a text description and an image of a single subject or images of multiple subjects, FastComposer encodes the image as a series of subject embeddings used to augment the prompt embeddings generated from textual input. Building upon FastComposer's methodology, we use delayed subject conditioning controlled by a hyperparameter, denoted as α, which influences when in the inference process we first introduce the reference image conditioning. Namely, we denoise conditioned on prompt embeddings at timesteps $t \leq \alpha T$, and prompt embeddings augmented with per-token subject embeddings at timesteps $t > \alpha T$, all on samples from ControlNet guidance. This time-dependent noise prediction model is defined as:

$$\epsilon_t = \begin{cases} \epsilon_\theta(z_t, t, c) & \text{if } t > \alpha T \\ \epsilon_\theta(z_t, t, c') & \text{if } t \leq \alpha T \end{cases} \tag{1}$$

where c is the original text embedding and c' is the text embedding augmented with the input image embedding.

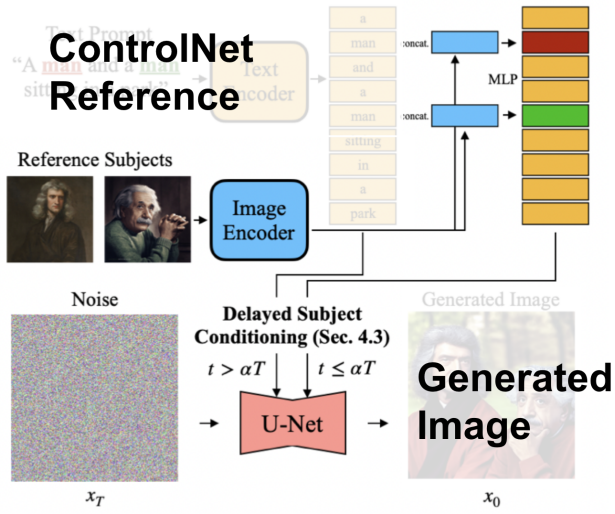


Figure 1: Model Architecture: Controlled FastComposer swaps out FastComposer's text-conditioned denoiser with ControlNet's text-conditioned denoiser.

Experiments. We investigate the Controlled FastComposer model under two settings: single-subject and multi-subject generation. In single-subject experiments, we employ a set of 100 faces from the Celeb-A dataset [5] combined with a single guiding image featuring a person dancing for ControlNet. We feed this image into ControlNet using PoseNet generations to infer the human pose. In multi-subject experiments, we adopt a similar approach by using a guiding image with multiple figures for our ControlNet input. However, we focus on a few images of faces with clearly distinct features, so we can assess multi-subject generation of distinct features.

To assess the role of delayed conditioning in Controlled FastComposer, we explore a range for α from 0 to 1, with increments of 0.1. The following boundary cases are particularly significant:

- When $\alpha = 0$, the model aligns with a purely text-based conditioning model, resembling the standard ControlNet without reference images.
- When $\alpha = 1$, the model operates without delayed conditioning, potentially leading to overfitting on reference images while neglecting the text prompt.

We used 50 inference steps and a classifier guidance scale of 9.

Evaluation. We evaluate the model on two fronts: its ability to maintain the identity of reference images and its adherence to the given text prompt.

- To evaluate identity preservation, we examine the L_2 norm (or Euclidean distance) between FaceNet embeddings [6] extracted from the reference and output images through Multi-Task Cascaded Convolutional Neural Networks (MTCNN). This is a similarity metric defined as:

$$L_2 \text{ Score} = \|E_{ref} - E_{out}\|_2 , \tag{2}$$

  where E_{ref} is the FaceNet embedding for the reference image and E_{out} is the FaceNet embedding for the output image.

- To evaluate prompt-image similarity, we use the CLIP Score computed by OpenCLIP [4]. This is a text-to-image similarity metric defined as:

$$\text{CLIPScore}(I, C) = \max(\cos(E_I, E_C), 0), \tag{3}$$

  where E_I is the visual CLIP embedding for an image I and E_C is the textual CLIP embedding for a caption C. This score, ranging between 0 and 1, captures the cosine similarity between the visual and textual embeddings, with higher values indicating a stronger correlation between the image and its associated caption.

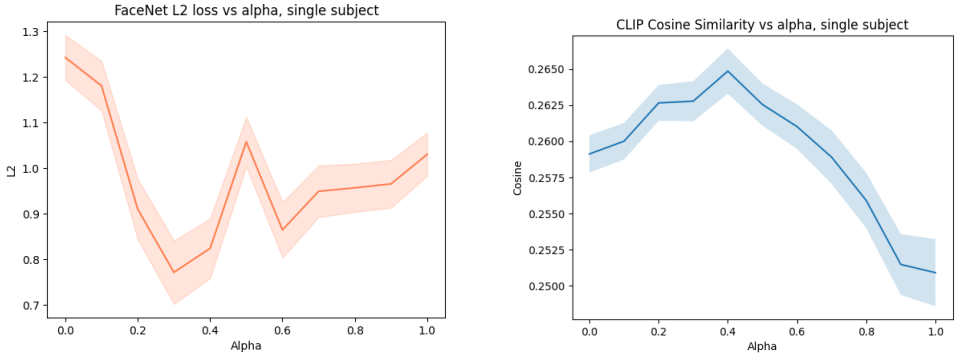
For each value of α, we report the mean scores across the evaluation datasets.

4 Results

Figure 2 shows the FaceNet L_2 scores and CLIP scores for single-subject generation. We can observe that introducing a certain degree of delay in conditioning demonstrates an advantage over both the baselines of ControlNet ($\alpha = 0$) and undelayed conditioning ($\alpha = 1$). For single-subject generation, α between 0.2 and 0.4 achieves a balance between prompt consistency and identity preservation. The specific values are found in Table 1, with the best score for FaceNet at $\alpha = 0.3$ and the best score for CLIP at $\alpha = 0.4$.

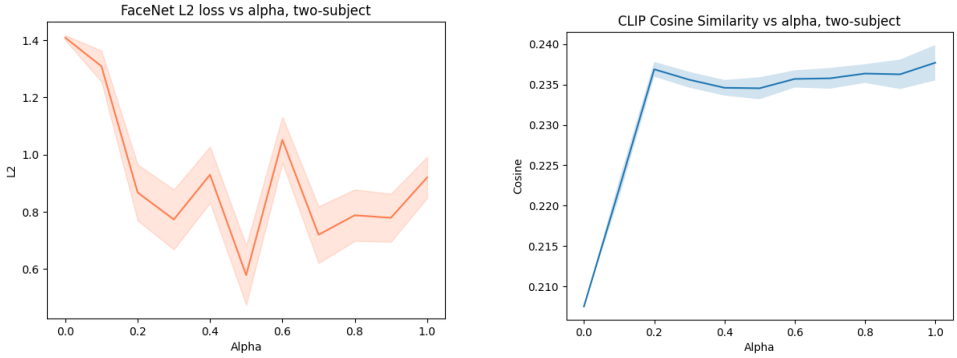
Figure 3 presents the same scores for multiple subject generation. We can observe that the optimal value of α at $\alpha = 0.5$ achieves lower identity preservation loss when compared to the reference images, but achieves comparable CLIP scores to α values ranging from 0.2 to 1. Across single and multiple subject generation in Table 1, without any reference images, the ControlNet alone struggles more with multiple subject generation compared to single-subject generation, but interestingly enough, the undelayed conditioning does not experience the same decrease in performance. On the other hand, all values of α experience a decrease in CLIP score when moving to multiple subject generation.

Figures 4 and 5 showcase examples of single-subject and multi-subject generations using the Controlled FastComposer architecture. Notably, Figure 5 allows for distinct feature generation even with the ControlNet architecture and avoids the problem of identity blending discussed in Section 2.



(a) FaceNet L_2 Score. (b) CLIPScore

Figure 2: Evaluations for Single-Subject Generation.



(a) FaceNet L_2 Score. (b) CLIPScore.

Figure 3: Evaluations for Multiple-Subject Generation.

Single-Subject Generation	L_2 **Score**	**CLIPScore**
Text-to-image ($\alpha = 0$, ControlNet)	1.242 ± 0.049	0.259 ± 0.001
Controlled FastComposer ($\alpha = 0.3$)	$\mathbf{0.771 \pm 0.069}$	0.263 ± 0.001
Controlled FastComposer ($\alpha = 0.4$)	0.824 ± 0.066	$\mathbf{0.264 \pm 0.002}$
No delayed conditioning ($\alpha = 1$)	1.030 ± 0.047	0.251 ± 0.002

Multi-Subject Generation	L_2 **Score**	**CLIPScore**
Text-to-image ($\alpha = 0$, ControlNet)	1.408 ± 0.008	0.208 ± 0.000
Controlled FastComposer ($\alpha = 0.5$)	$\mathbf{0.579} \pm 0.104$	0.235 ± 0.001
No delayed conditioning ($\alpha = 1$)	0.920 ± 0.072	$\mathbf{0.237 \pm 0.002}$

Table 1: L_2 scores, for identity preservation, and CLIP scores, for prompt-image similarity, across optimal values of delay conditioning α in both single-subject and two-subject experiments. For L_2 scores, lower values are better; for CLIP scores, higher values are better.

Figure 4: Reference Image, ControlNet Image, PoseNet Output, and Generated Image for Single-Subject Generation.

Figure 5: Reference Images, ControlNet Image, PoseNet Output, and Generated Image for Multi-Subject Generation.

5 Conclusion and Future Research

In this paper, we propose Controlled FastComposer, an integration of FastComposer with ControlNet, aimed at addressing the limitations of spatial control in text-to-image generation. Our experiments demonstrate that introducing delayed subject conditioning in Controlled FastComposer allows for a more balanced trade-off between identity preservation and prompt-image similarity, as evidenced by improved FaceNet L_2 scores and CLIP scores. Controlled FastComposer surpasses the performance observed with ControlNet alone or with undelayed conditioning.

To extend the scope of Controlled FastComposer, several avenues for future research can be explored. For example, we can expand the model's ability in the context of multi-subject generation to handle scenarios involving more two people in the input images. Moreover, we can explore a broader range of prompts to enhance the model's adaptability to different creative requirements. Finally, further investigation into fine-tuning strategies could refine the model's performance for specific use cases, and conducting memory usage evaluations could ensure the model's practicality in resource-intensive environments to allow for potential deployment in real-world applications.

6 Limitations

While Controlled FastComposer shows promising advancements in text-to-image generation, the model exhibits sensitivity to prompt selection, with the effectiveness of the delayed conditioning parameter α affected by choosing an appropriate prompt. In addition, the model's performance is constrained by the size of the dataset used for training and the availability of computational resources. The current limitations in the dataset may impact the model's ability to generalize across diverse scenarios, and resource constraints may influence the model's scalability and efficiency.

Author Contributions

All team members were fully involved in the writing of the report, editing of the report, and qualitative and quantitative evaluation of the results.

Alex Hu. Built the model and initial version of the inference code.

Sean Li. Wrote the model architecture and contributed to the experimental design (CLIP textual similarity).

Emily Liu. Spearheaded the experimental design. Built the inference and evaluation code for single and multiple-subject generation. Created figures. Contributed to writing the experiments, evaluation, and results.

Sarah Zhang. Defined the metrics in the evaluation. Wrote the abstract, introduction, related work, conclusion and future research, and limitations. Contributed to writing the model architecture, experiments, evaluation, and results.

Acknowledgements

We express sincere appreciation to Professor Song Han for introducing us to this research area and providing valuable insights into efficient deep learning techniques. We also express gratitude to Guangxuan Xiao for providing invaluable feedback on the proposal of the paper. Furthermore, we would like to express our thanks to Han Cai and Ji Lin for their unwavering support and assistance throughout the semester.

References

[1] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869, 2023.

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *CoRR*, 2021.

[3] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[4] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. URL `http://arxiv.org/abs/1503.03832`.

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

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.