Enhanced ControlNet: Refining Multimodal Image Generation with Integrated Text and Sketch Guidance

Cherukupalli Sai Malini Mouktika
IIT Hyderabad
ai21btech11007@iith.ac.in

Beeram Sandya
IIT Hyderabad
cs21btech11006@iith.ac.in

Challa Akshay Santoshi
IIT Hyderabad
cs21btech11012@iith.ac.in

Cheekatla Hema Sri
IIT Hyderabad
cs21btech11013@iith.ac.in

Abstract

Stable Diffusion has emerged as a powerful generative model for high-quality image synthesis, enabling applications such as text-to-image generation, inpainting, and style transfer. However, precise control over image generation remains a challenge, particularly when integrating structural constraints like edge maps, depth maps, or human poses. ***ControlNet*** *addresses this issue by conditioning Stable Diffusion on additional control signals while preserving the original model's generative power. In this work, we train a* ***ControlNet*** *on a selected dataset and evaluate its performance in terms of prompt fidelity, condition adherence, and image quality. The training process involves fine-tuning a set of trainable layers while keeping the original* ***UNet*** *of Stable Diffusion frozen, ensuring efficient learning without excessive computational cost. We document the challenges faced during training, such as dataset preparation, hyperparameter tuning, and computational constraints. This project provides insights into the practical implementation of ControlNet, highlighting its advantages and limitations. Additionally, we evaluate its performance across multiple conditioning types and propose optimizations to reduce inference latency.This work contributes to improving controllability in diffusion-based image synthesis, making it more applicable to real-world creative applications.*

1. Introduction

Image synthesis has witnessed significant advancements with the development of deep generative models. Among these, Stable Diffusion has emerged as a leading method for generating high-quality images from text prompts, enabling applications such as text-to-image generation, inpainting, and style transfer. This has opened new possibilities in creative industries, digital content generation, and interactive AI-driven design. However, despite its success in generating realistic and visually compelling images, Stable Diffusion lacks precise control over image structure and composition. Generating images that align perfectly with a user's mental concept remains a challenge, often requiring multiple iterations of prompt engineering.

One of the main limitations of text-to-image models like Stable Diffusion is their inability to enforce spatial constraints effectively. While users can guide the model with textual descriptions, they lack fine-grained control over object placement, structure, and specific image attributes. This issue becomes particularly evident in scenarios where structural information—such as edge maps, depth maps, human poses, or segmentation maps—is crucial to preserving an image's intended composition. These structural constraints are essential in applications such as medical imaging, architectural visualization, human pose-guided synthesis, and sketch-based image generation.

To address these limitations, ControlNet has been introduced as a neural network architecture that extends Stable Diffusion by incorporating additional conditioning inputs. ControlNet enables spatially precise image generation by conditioning the model on external control signals while maintaining the original generative power of Stable Diffusion. By doing so, it provides users with greater flexibility in directing the diffusion process, ensuring that generated images adhere more closely to their desired structure and composition.

This project aims to conduct a thorough investigation into the current state of ControlNet-based image generation, with a focus on how text prompts and spatial cues (such as sketches) can be more effectively combined to drive high-quality image synthesis.

2. Literature Survey

Recent advances in generative modeling have led to remarkable progress in text-to-image and image-to-image synthesis. However, a persistent challenge remains: integrating multimodal control (e.g., combining natural language with spatial guidance such as sketches or bounding boxes) in a seamless and efficient manner. In this section, we review the evolution of methods—from early diffusion models and autoregressive approaches like DALL·E, to specialized control frameworks such as ControlNet and instruction-guided editing models like InstructPix2Pix—and discuss their limitations and how successive work has addressed these issues.

2.1. Denoising Diffusion Probabilistic Models (DDPMs):

Denoising Diffusion Probabilistic Models[2] formulate image generation as a gradual denoising process, in which a neural network is trained to reverse a fixed forward process that progressively corrupts an image with Gaussian noise. This framework produces high-quality, photorealistic images and offers a principled, likelihood-based approach. However, DDPMs are originally designed for unconditional or basic text-conditional generation and do not provide a mechanism for incorporating explicit spatial guidance (e.g., sketches or bounding boxes). Moreover, the iterative sampling process can be computationally intensive, and fine control over local details remains challenging.

2.2. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2):

Building on the power of contrastive models, DALL·E 2[5] adopts a hierarchical framework that decouples text-to-image synthesis into two stages. In the first stage, a prior network generates a CLIP [4] image embedding from a given text caption; in the second stage, a decoder (often based on diffusion models) synthesizes an image conditioned on that embedding. This separation enables better semantic alignment and diversity. However, while this approach effectively leverages CLIP's rich joint representation of text and images, it primarily focuses on the global semantic content of the generated image and does not inherently address spatial constraints, leaving limited options for precise layout or structure control.

2.3. High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion):

Stable Diffusion[6] further advances the efficiency of diffusion-based generation by performing the diffusion process in a lower-dimensional latent space, which allows for high-resolution image synthesis at a fraction of the computational cost. Although this method achieves impressive photorealism and speed, its conditioning is typically limited to text alone, and it does not incorporate additional spatial signals. As a result, while Stable Diffusion excels at generating diverse and high-quality images, it struggles to integrate explicit spatial guidance—an essential requirement for tasks where control over the layout (such as sketches or bounding boxes) is needed.

2.4. Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet):

ControlNet[8] directly addresses the limitation of text-only conditioning by introducing a modular control branch that accepts additional spatial inputs (e.g., sketches, edge maps, or bounding boxes) alongside text prompts. In ControlNet, the pretrained diffusion model's weights are frozen to preserve its high-quality image synthesis capabilities, while the lightweight control branch (often initialized with zero convolution layers) learns to inject spatial guidance into the network. This design significantly improves controllability without sacrificing photorealism. Despite its advantages, ControlNet remains sensitive to the quality and alignment of the control signal and may require careful tuning when handling complex or multiple modalities of guidance.

2.5. InstructPix2Pix: Learning to Follow Image Editing Instructions:

InstructPix2Pix[1] adapts diffusion models to perform instruction-guided image editing. Rather than generating images from scratch, this model fine-tunes a pretrained diffusion network so that it can modify an input image based on a natural language instruction (e.g., "make the sky purple"). By conditioning on both the input image and the editing command, InstructPix2Pix demonstrates the feasibility of guided image manipulation. However, its focus is on editing rather than generation from combined text and spatial cues. Moreover, because the training data often relies on synthetic editing pairs, the model can sometimes struggle with ambiguous instructions or with maintaining the photorealism of the original image.

2.6. Sketch-Guided Image Generation (Sketch Your Own GAN, Chen et al., ICCV 2023):

Sketch-based image generation has been explored extensively, ranging from retrieval-based methods to deep generative models. GAN Sketching [7] introduces a method to modify a pre-trained GAN using user-provided sketches, where the model's weights are adjusted to match the provided shapes while preserving realism. Their approach utilizes a cross-domain adversarial loss, where a pre-trained image-to-sketch transformation network helps guide the adaptation process. While GAN Sketching enables model customization, its reliance on GAN fine-tuning introduces challenges such as training cost, overfitting, and limited diversity in generated images. Our work builds on the idea of sketch-conditioned image generation but extends it to diffu-

sion models through ControlNet, which enables sketch-to-image generation without requiring model fine-tuning. Unlike GAN Sketching, our method preserves the pre-trained diffusion model, conditioning the synthesis process dynamically to enhance the structure and details of rough inputs.

2.7. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback

ControlNet++ [3] enhance the controllability of text-to-image diffusion models, existing efforts like ControlNet incorporated image-based conditional controls. Traditional diffusion models generate images based on the text prompts but often lack the fine-grained control over the structure of generated images. ControlNet++ addresses thegap between the high level semantic guidance (text prompts) and low level conditional control. The key contributions of ControlNet++ include enhanced control mechanisms through improved integration of external conditioning information for better structure preservation, an efficient training strategy that reduces computational overhead while maintaining high-quality image synthesis.

Diffusion models, particularly DDPMs and Stable Diffusion has gone substantial advancements in generation of images. In text-to-image synthesis, diffusion models integrate cross-attention mechanisms between UNet-based denoisers and text embeddings from pre-trained language models like CLIP and T5, enabling reasonable text-to-image generation. Despite the astonishing capabilities of diffusion models, language is a sparse and highly semantic representation, unsuitable for describing dense, low-semantic images.Furthermore, existing methods still struggle to understand detailed text prompts, posing a severe challenge to the controllable generation.

To achieve conditional control in pre-trained text-to-image diffusion models, ControlNet introduce additional trainable modules for guided image generation. However, they still lacks a clear approach for enchancing the controllability under various controls. Furthermore, existing works implicitly learn controllability by the denoising process of diffusion models, while our ControlNet++ achieves this in an explicit cycle-consistency manner.

The reward model evaluates how well generative model outputs align with human expectations, improving controllability. Initially used in NLP via RLHF, it has expanded to vision tasks for enhancing image quality. However, image quality is subjective and requires specialized datasets and trained reward models. Instead of focusing on subjective preferences, this work emphasizes precise controllability, using AI feedback as a more cost-effective alternative to human feedback.

In the ControlNet++ model, to enchance the controllability as the consistency feedback between input conditions and generated images, the outcome is quantified through the discrimination reward models as Reward Consistency Loss. Diffusion training loss is also employed so that the original image generation capability is not compromised. Additionally, instead of randomly sampling from noise like in diffusion models, we add noise to the training images, thereby explicitly disturbing the consistency between the diffusion inputs and their conditional controls.

$$L_{\text{total}} = L_{\text{train}} + \lambda \cdot L_{\text{reward}} \tag{1}$$

where λ is a hyperparameter to adjust the weight of reward loss.

2.8. Synthesis and Our Approach:

The progression from DDPMs to hierarchical methods like DALL·E 2 and Stable Diffusion has yielded state-of-the-art quality in text-to-image generation, yet these models traditionally lack mechanisms to integrate fine-grained spatial guidance. ControlNet represents a key step toward addressing this gap by allowing diffusion models to accept both text and image conditions, thereby significantly improving controllability. Our project builds on this insight by implementing a ControlNet-like architecture that combines text prompts with spatial cues (e.g., sketches or bounding boxes) to enable more versatile multimodal image generation.

Figure 1. Model used: stable-diffusion-v1-5/stable-diffusion-v1-5

Figure 2. Model used: sd-legacy/stable-diffusion-v1-5

Figure 3. Model used: CompVis/stable-diffusion-v1-4

Figure 4. Model used: runwayml/stable-diffusion-v1-5

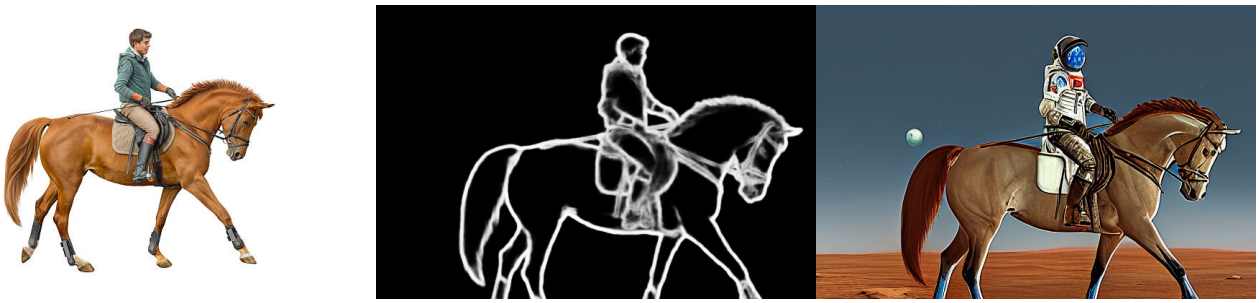

Figure 5. orignial-hed-generated images

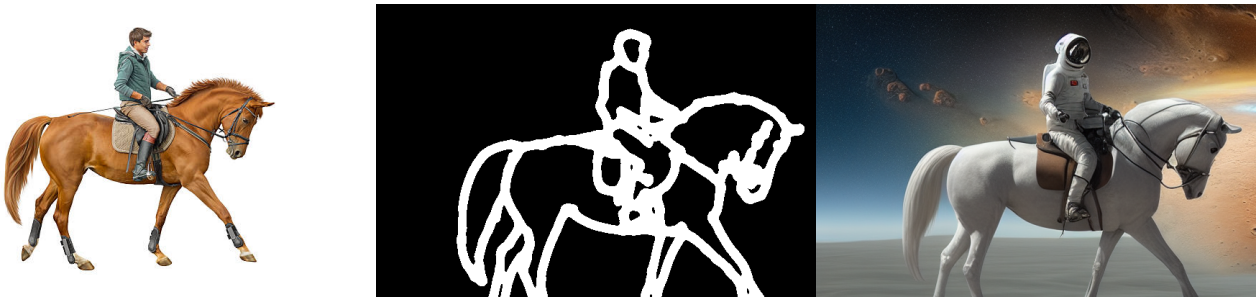

Figure 6. original-scribble-generated images

Figure 7. original-depth-generated images

Figure 8. original-openpose-generated

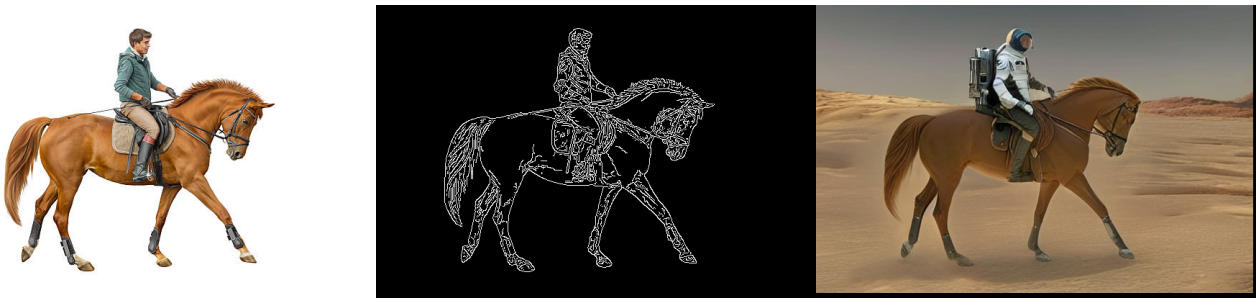

Figure 9. original-canny-generated images

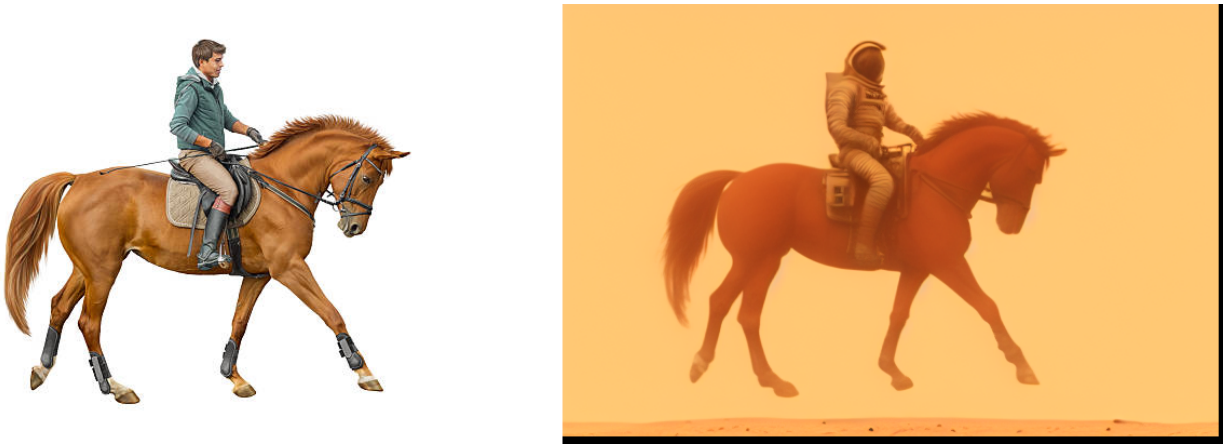

Figure 10. Orignal and generated images (instructPix2Pix)

3. Replication of Literature Results

3.1. Diffusion Models

In this section, we analyze images generated by four diffusion models using the prompt "a photo of an astronaut riding a horse on Mars." The resulting images are evaluated for quality, style, and adherence to the prompt. The Stable Diffusion v1 models demonstrate strong text-to-image generation capabilities, with v1.5 checkpoints (Figure 1, Figure 2, Figure 4) outperforming v1.4 (Figure 3) in quality and detail.

The official Stable Diffusion v1.5 checkpoint was trained for 595,000 steps at 512x512 resolution on the laion-aesthetics v2 5+ dataset, with 10% text-conditioning dropping for improved classifier-free guidance. The Stable Diffusion v1.4 was trained for 225,000 steps at 512x512 on laion-aesthetics v2 5+, with 10% text-conditioning dropping. It's an earlier iteration than v1.5, resumed from v1-2.

All models successfully interpreted the surreal prompt, placing an astronaut on a horse in a Mars-like environment. Figure 3 exhibits more artifacts or less realistic proportions compared to v1.5 models. Its aesthetic quality is still high, but it lacks the refinement seen in Figure 1 and Figure 2, potentially showing blurrier textures or less accurate lighting. Figure 1 and Figure 4 (official and RunwayML v1.5) likely offer the most polished outputs, while Figure 2 (legacy v1.5) introduces a more dynamic, stylized interpretation.

3.2. ControlNet Models

In this section, we are showing how different conditioning methods influence the output of text-to-image generation models. Using Stable Diffusion v1.5 as the base model, we integrated various ControlNet variants—each using a different type of conditioning image (canny edges, depth map, scribble, HED, and openpose)—to generate images from the prompt "a realistic image of an astronaut riding a horse on mars". Stable Diffusion, a latent diffusion model, has become one of the most popular frameworks due to its open-source availability and efficiency. ControlNet extends this model by incorporating an extra conditioning signal (such as canny edges, depth maps, etc.) that guides the generation process, allowing for finer control over image composition.

1. **Canny Edge Conditioning**: Canny image is a monochrome image with white edges on a black background. The figure 9 shows the actual image, corresponding canny image and then corresponding ControlNet generated images respectively. The output preserves the crisp outlines, leading to a well-defined structure. Canny images offers clear boundaries and enhances structural details. However, it may sometimes oversimplify texture details.
2. **Depth Map Conditioning** Depth image is a grayscale image with black representing deep areas and white representing shallow areas. The figure 7 shows the actual image, corresponding depth map and the corresponding ControlNet generated images respectively. The depth conditioning helps maintain spatial consistency, especially in the background and overall layout. Depth maps enhances spatial depth and can improve background coherence, which is beneficial for scenes with significant depth like "Mars.
3. **Scribble and HED Conditioning** Scribble image is a hand-drawn monochrome image with white outlines on a black background. HED image is a monochrome image with white soft edges on a black background. 6 and 5 shows repective original, conditioned and generated images of scribble conditioned ControlNet model and HED conditioned ControlNet model. Scribble Conditioned image variant introduces creative textures and stylization, adding an artistic feel to the generated image. In HED conditioned image the soft edge information improves the overall smoothness of details and can enhance the photo-realism. Both provide edge information but with different characteristics. Scribble-based conditioning tends to introduce an artistic flair, while HED delivers smoother transitions and can support more photo-realistic outputs.
4. **Openpose Conditioning** 8 shows the original, corresponding openpose image and corresponding ControlNet generated image. This approach emphasizes human pose, which can improve the depiction of the rider's posture and dynamic movement.It is particularly useful for maintaining accurate human pose. In this case, it helps improve the depiction of the rider's posture and movement, contributing to a more dynamic composition.

3.3. InstructPix2Pix

InstructPix2Pix is a deep learning model designed for instruction-based image editing. It enables users to modify images using natural language prompts, making it a powerful tool for creative and interactive image manipulation. A text prompt ”a realistic image of an astronaut riding a horse on mars” and a reference is given as input. In 10 the left image is given as input and the right image is the generated image.

References

[1] Tim Brooks, Aleksander Holynski, Alexei A. Efros, and Berkeley University of California. Instructpix2pix: Learning to follow image editing instructions. 2023. 2

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020. 2

[3] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. 2024. 3

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2021. 2

[5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2022. 2

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[7] Sheng-Yu Wang, David Bau, Jun-Yan Zhu, Carnegie Mellon University, and MIT CSAIL. Sketch your own gan. 2021. 2

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala Stanford University. Adding conditional control to text-to-image diffusion models, 2023. 2