

KnobGen: Controlling the Sophistication of Artwork in Sketch-Based Diffusion Models

Supplementary Material

Appendix

In this appendix, we provide additional details about the model architecture and supplementary results that further demonstrate the robustness of our approach. These sections aim to provide a deeper understanding of the technical components and showcase more comprehensive comparisons.

- Appendix 1: explain training and evaluation setup.
- Appendix 2: expands on the details of the CGC module.
- Appendix 3: explains our conducted user study.
- Appendix 4: provides more qualitative results.

1. Setup

Dataset: We utilized the MultiGen-20M dataset, as introduced by [10], to train and evaluate our model. The dataset offers various conditions, making it a suitable choice for our approach. We selected 20,000 images for training, focusing specifically on those with the Holistically-nested Edge Detection (HED) [14] condition. However, we modified the KnobGen condition by applying a thresholding technique, where pixels below a threshold value of 50 were set to zero, and those above were set to one. This threshold value was chosen through simple visual comparisons of several samples using different thresholds, allowing us to identify the most effective value. This modification essentially transforms the HED condition into a sketch. For evaluation, we curated two distinct sets of images. The first evaluation set consisted of 500 randomly selected samples, which are similar to a sketch drawn by a seasoned artist (we followed the thresholding technique for this part), allowing us to measure our model’s effectiveness in professional settings. To further test the robustness and adaptability of our approach, we compiled a second evaluation set of 100 hand-drawn images created by non-professional individuals. This diverse testing set enabled us to demonstrate the model’s ability to generalize across a broad spectrum of users, ensuring it can handle both professionally designed and amateur drawings with high robustness.

Baselines: In this work, we evaluate the performance of our proposed model against several state-of-the-art (SOTA) diffusion-based models. Specifically, we conduct both qualitative and quantitative comparisons with prominent models such as ControlNet [15], T2I-Adapter [8], AnimateDiff [1], UniControl [10], and ControlNet++ [6]. These models have achieved significant advances in fine-grained control of image generation by incorporating sketch-based condi-

tions into the diffusion process. Since AnimateDiff is a video-based DM, we only use the first frame of the generated video by it as the comparison point.

Evaluation: We perform qualitative and quantitative evaluation. In the qualitative evaluation, we compare our model’s performance across different scenarios of varying input conditions and complexities. For quantitative evaluation, we utilize several metrics to assess the quality of the generated images. First, we calculate the Fréchet Inception Distance (FID) [2, 4], which measures the similarity between generated and natural images using a pre-trained InceptionV3 model [13]. Lower FID values indicate better generation quality; we used [3] implementation for our evaluation, which used the default pre-trained InceptionV3 model available in Pytorch [9]. To evaluate the alignment between the generated images and the text prompts, we use CLIP [11], specifically the pre-trained DetailCLIP model [7] with a Vision Transformer (ViT-B/16) backbone. Higher CLIP scores signify better alignment between the generated images and their corresponding prompts. Finally, we assess the realism and aesthetic quality of the generated images using the metric proposed by [5], where higher scores reflect more visually appealing images.

Implementation Details: Our proposed *KnobGen* framework is built on top of Stable Diffusion v1.5 [12], with the original parameters kept frozen throughout training. For the Fine-Grained Controller (FGC) module, we employed two different pre-trained models to demonstrate the flexibility and effectiveness of our approach across multiple setups. Specifically, we integrated ControlNet [15] and T2I-Adapter [8], both of which had their parameters frozen and were not updated during training. The architecture and integration of these components are illustrated in Figure ?? . We trained the CGC module for a total of 2000 epochs using 16 A100 GPUs. During the initial 1500 epochs, we employed the modulator mechanism, as described in Section ?? , with a learning rate of $1e - 5$. In the final 500 epochs, we fine-tuned the CGC model with a reduced learning rate of $1e - 6$ to ensure robustness and to improve the quality of the generated images.

2. Model architecture

The CFC module plays a critical role in our model by integrating and aligning visual and textual information for effective image generation. The primary goal of the CFC is

to ensure that features derived from both the input sketch image and the text prompt are jointly fused, allowing the model to generate more contextually relevant and visually coherent outputs. The CFC module has around 100M trainable parameters.

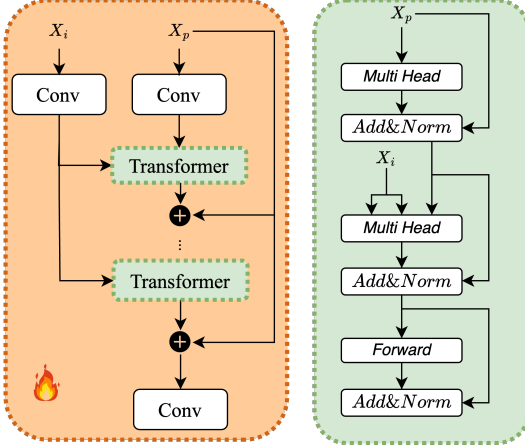


Figure 1. **Overview of the Cross-Feature Conditioning (CFC) module.** The module integrates visual and textual features through a series of transformer blocks with cross-attention. In the diagram, X_i represents the encoded image features from a sketch, while X_p denotes the encoded text prompt. The CFC module conditions the text features based on the image input, allowing for fine-grained control and alignment between visual and textual inputs during the image generation process.

To achieve this, we designed the CFC module using a transformer-based architecture that leverages cross-attention between image and text features; Figure 1 shows the CFC overview. Below, we explain the architecture and functionality in detail:

Architecture: The CFC module is composed of three key components: convolutional layers for feature transformation, transformer layers for cross-attention, and fully connected layers for output projection. The module takes two inputs—visual features (encoded input image) and text features (encoded text prompt)—and processes them jointly to output contextually conditioned text features.

- **1D Convolutional Layers:** The input to the CFC module consists of two tensors: an encoded image tensor $x_i \in \mathbb{R}^{\text{batch} \times 256 \times 1024}$, which comes from CLIP image encoder, and an encoded text tensor $x_p \in \mathbb{R}^{\text{batch} \times 77 \times 768}$, which comes from text encoder of CLIP like all the prompt conditioned DM. We then pass these embeddings through 1D convolutional layers to project the input channels (1024 for images and 768 for text) into a common hidden dimension of 1024 channels. This transformation ensures that both modalities can be effectively combined in the cross-attention mechanism.

- **Transformer Layers for Cross-Attention:** The core of the CFC module lies in its eight layers of transformers that perform cross-attention. These layers allow the model to fuse information from both the image and text features. Specifically, the image tensor serves as the memory input for the transformer, while the text tensor undergoes cross-attention, attending to the visual information. This design enables the model to enhance text-based features by conditioning them on the spatial and structural content of the image. The resulting enriched text features better capture the contextual relevance of the image, leading to more semantically meaningful generation.
- **Fully Connected Layers:** After passing through the transformer layers, the output text tensor is reduced back to its original sequence length (77 tokens) and further processed through two fully connected layers. These layers refine the text features, ensuring that the final output has the desired dimensionality (batch, 77, 768) and captures the relevant information for conditioning the image generation process.

Reasoning Behind the Design: The CFC module is specifically designed to address the need for strong alignment between visual and textual inputs during image generation. By using a cross-attention mechanism, the module ensures that the text features are not treated independently of the visual content, but rather, are conditioned on the image’s features as well. This approach is particularly useful when fine-grained control is needed to generate images that aligns to both the textual description and visual input, making it highly effective in scenarios where accurate text-to-image alignment is crucial. Additionally, the use of pre-trained models ensures that the model benefits from robust initial feature extraction which further improves generation quality as a result.

3. More Quantitative Result: User Study

In this section we provide an additional quantitative evaluation which is our conducted user-study experiment.

We conducted a user study to gain deeper insights into the perceived quality and usability of KnobGen, compared to existing state-of-the-art baselines. This study involved 100 participants, who were asked to evaluate a set of 10 images selected randomly from a set of 50 generated images with different complexities across *three* key dimensions: *sketch alignment*, *prompt alignment*, and *aesthetic quality*. Sketch alignment measures how well the generated image adheres to the spatial and structural details of the input sketch, while prompt alignment assesses the consistency between the generated image and the provided textual prompt. Aesthetic score captures the overall visual appeal of the generated images. Participants rated each dimension on a scale from 0 to 10, with higher scores indicating better

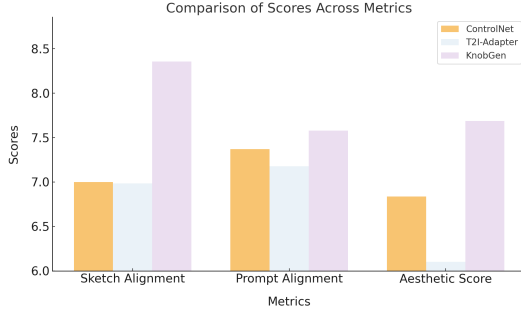


Figure 2. **Comparison of model performance in our user study:** Our user study includes three metrics sketch alignment, prompt alignment, and aesthetic score. While all models perform similarly in prompt alignment, KnobGen significantly outperforms ControlNet and T2I-Adapter in sketch alignment and aesthetic score, demonstrating its superior capability in handling sketch precision and generating visually appealing outputs.

performance. While models are equally performing in the prompt alignment as shown in Figure 2, KnobGen exhibits a distinct edge over handling diverse set of sketches while maintaining a visually appealing generation.

4. More Qualitative Result

This section contains more qualitative results to complement the evaluations presented in the main paper. We provide visual examples of different use cases, including scenarios involving amateur and professional sketches.

4.1. Inference Knob Mechanism For Baselines

One of the important ablation studies was to evaluate the performance of fine-grained controller models, such as the T2I-Adapter, when they utilize our Knob mechanism. This ablation study was particularly performed to demonstrate the effectiveness of our proposed CGC module.

Models such as T2I-Adapter are traditionally designed for precise, detail-oriented image generation but lack the flexibility to accommodate broader, more abstract inputs like rough sketches or varying user skills. To explore this issue, we integrated the Knob system into the T2I-Adapter model **without our CGC module**.

Figure 3 showed that while the T2I-Adapter performs exceptionally well in generating high-fidelity images from professional-grade inputs, it struggles to maintain this quality when dealing with rougher or less detailed sketches. This limitation arises from the absence of a Macro Pathway in the T2I-Adapter’s architecture, which makes the model overly reliant on precise input details. Without the ability to capture broader, high-level semantic information through a coarse-grained approach, the model becomes highly sensitive to adjustments made by the Knob mechanism. As a result, T2I-Adapter fails to deliver consistently good results

across a diverse range of users, particularly those providing amateur or less-defined sketches. Additionally, we observed that after a certain point, increasing the Knob value no longer meaningfully affects the generation output. This suggests that the sketch condition in T2I-Adapter influences the generation primarily in the early denoising steps, with diminishing effects in the later steps. However, further investigation of this behavior is outside the scope of this study.

While the Knob system is designed to balance coarse and fine-grained controls dynamically, the lack of a dedicated coarse-grained module in T2I-Adapter causes the model to lose spatial coherence when we apply our Knob mechanism for it, especially when the knob has low value. This issue became particularly evident when trying to generate images based on prompt only, as the model struggled to infer the missing spatial structure, leading to incoherent outputs.

In contrast, the KnobGen framework, including the CGC and FGC, demonstrated superior flexibility and performance. By incorporating both high-level abstractions and detailed refinements, KnobGen could adapt dynamically to the varying levels of detail in the input sketches. The CGC in KnobGen helps preserve the overall structure and semantics of the image, while the FGC ensures that fine details are accurately rendered.

4.2. More Qualitative Results

In this section, we present additional qualitative results to demonstrate the effectiveness and versatility of our proposed KnobGen framework further. Figure 4 showcases the model’s ability to handle a wide range of input sketches, from highly detailed professional-grade drawings to rough, amateur sketches.

References

- [1] Yuwei Guo and Ceyuan Yang. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1
- [2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1
- [3] Håkon Hukkelås. A Pytorch Implementation of the Fréchet Inception Distance (FID). <https://github.com/hukkelas/pytorch-frechet-inception-distance>, 2020. Version 1.0.0. 1
- [4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1
- [5] Junjie Ke, Keren Ye, and Jiahui Yu. Vila: Learning image aesthetics from user comments with vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10041–10051, 2023. 1

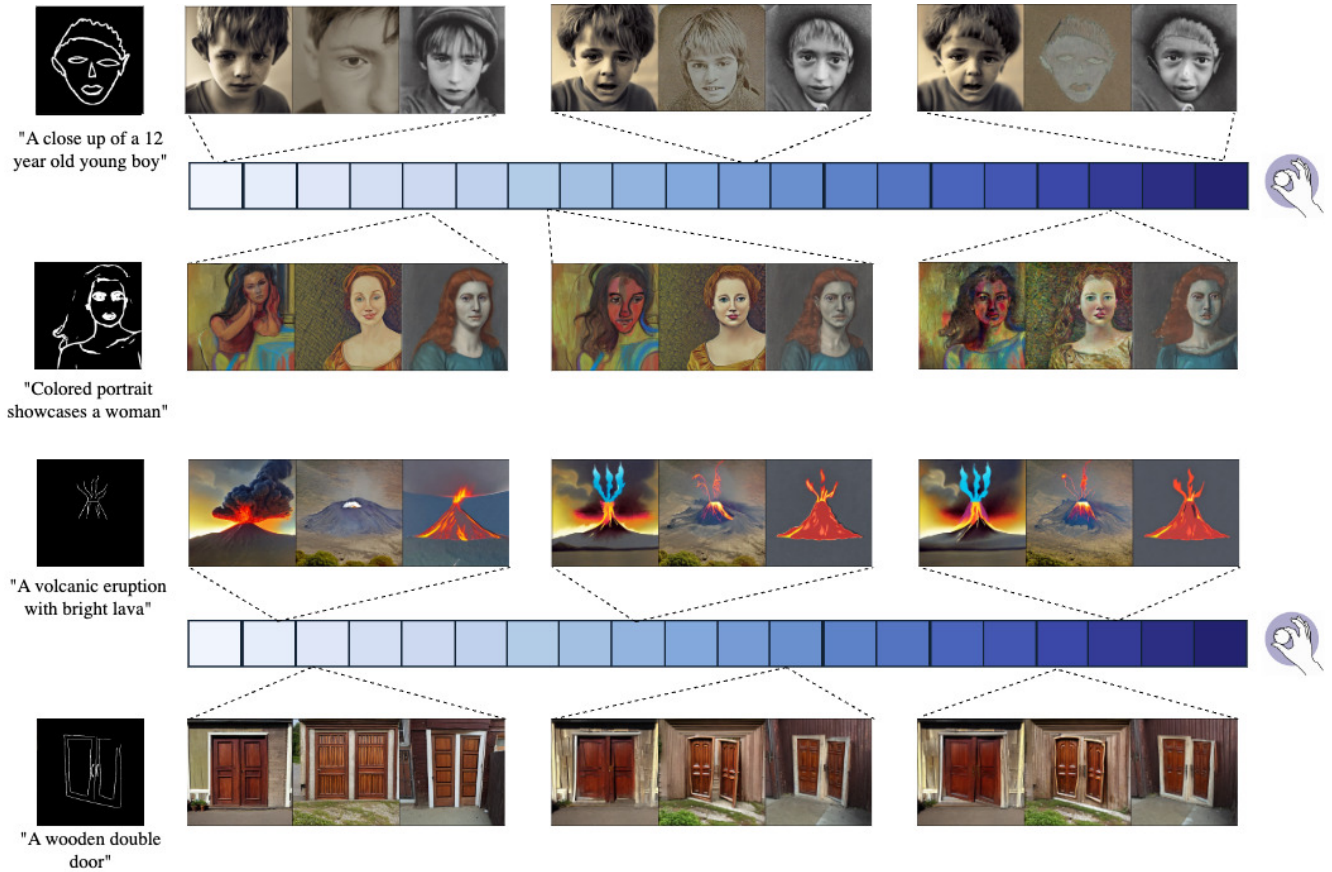
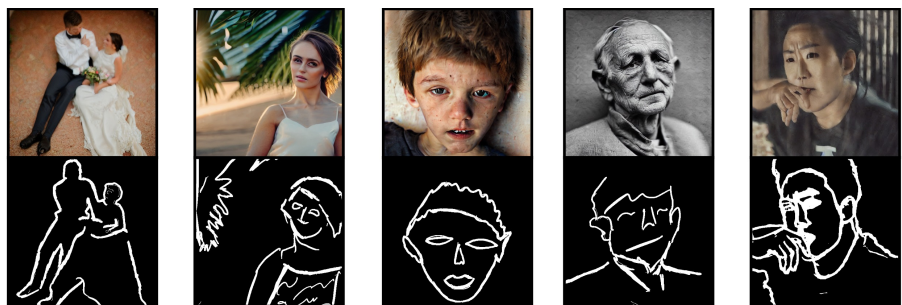


Figure 3. **Effect of the Knob mechanism on the fine-grained models (T2I-Adapter).** The image demonstrates how increasing the Knob value influences the generated output. While the T2I-Adapter performs well with precise, detailed sketches, it struggles with rougher sketches and fails to maintain spatial consistency as the Knob value increases. Beyond a certain threshold, the sketch has minimal impact on the final output, highlighting the model’s sensitivity to early-stage adjustments and its limitations in handling coarse-grained information.

- [6] Ming Li and Taojiannan Yang. Controlnet++: Improving conditional controls with efficient consistency feedback. *arXiv preprint arXiv:2404.07987*, 2024. 1
- [7] Amin Karimi Monsefi, Kishore Prakash Sailaja, Ali Alilooee, Ser-Nam Lim, and Rajiv Ramnath. Detailclip: Detail-oriented clip for fine-grained tasks. *arXiv preprint arXiv:2409.06809*, 2024. 1
- [8] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1
- [9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 1
- [10] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 1
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1
- [12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1
- [13] Christian Szegedy, Vincent Vanhoucke, and Sergey Ioffe. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1
- [14] Saining Xie and Zhuowen Tu. Holistically-nested edge detection, 2015. 1
- [15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1



A wedding picture
of a groom looking
at the bride

Portrait of a
vibrant young
lady next
to a palm tree

A closeup
of a 12 year
old young boy

Closeup photo
of a worn-
out old man

Song Seung-heon



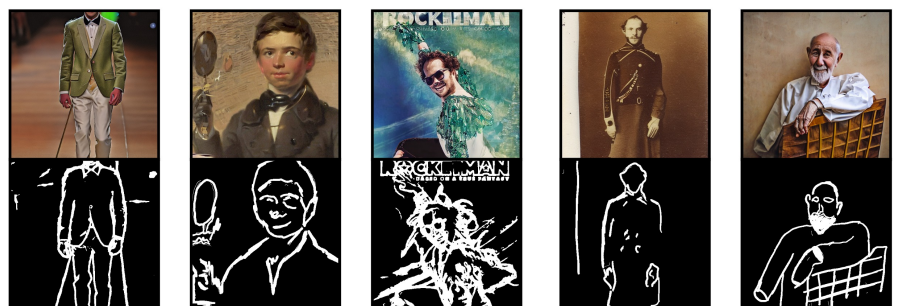
Girl in market
Dassa Benin

A young man with
long fringes

A woman in a
formal dress

Three people
next to
each other

Two children
building a tall tower



Men apparel

A classy man
holding a mirror,
UK 1840

Poster of
Rocketman

A Tsar soldier in a
uniform

A balding
old man



Tenor Andrea

Andrew Carnegie

Slim cut
striped silk

A female priest
laughing

Side portrait of a
lady wearing a
beautiful earring

Figure 4. More qualitative results on novice and professional-grade sketches