Test-time Conditional Text-to-Image Synthesis Using Diffusion Models (Supplementary Material)

Tripti Shukla[1,2*] Srikrishna Karanam[2] Balaji Vasan Srinivasan[2]
[1]Georgia Institute of Technology [2]Adobe Research, Bangalore, India
tshukla9@gatech.edu, {skaranam, balsrini}@adobe.com

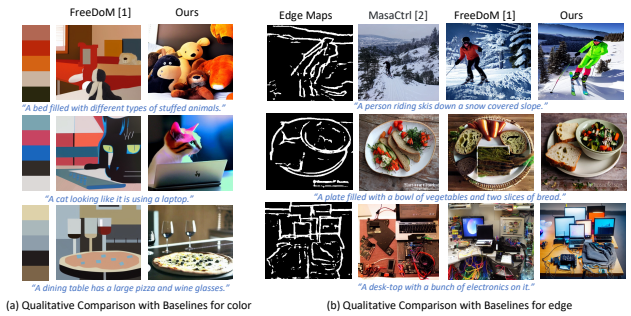

Figure 1. Qualitative comparison of our method with other training-free methods.

A. Additional Results

In Section A.1, we show additional qualitative results for our method against some state-of-the-art training-required methods like T2I-Adapter [3] and Control-Net [2]. In section B, we provide some additional implementation details about our method.

A.1. Qualitative Results

As depicted in Figure 2, TINTIN exhibits a superior capability in producing images that closely align with the color distribution inherent in the provided color palettes, surpassing alternative methodologies. Notably, the generated images by TINTIN demonstrate a heightened fidelity to the given prompt in comparison to the outputs from T2I-Adapter [3] and Control-Net [2]. In Figure 3, we present a visual comparison of images generated by our approach alongside those generated by the aforementioned methods, each conditioned on corresponding edge maps. Evidently, TINTIN-generated images exhibit a commendable adherence to the structural cues provided by the edge maps, outperforming alternative approaches in this regard. In Figure 1, we also provide qualitative comparison of TINTIN with some other training-free approaches like FreeDoM [5] and MasaCtrl [1] for both color and edge conditioning. We observe that our method not only performs better than some training-dependent approaches but is also superior to many training-free methods.

B. Implementation Details

In our approach, we leverage the pre-trained Stable Diffusion model [4], specifically version 1.4, which is also employed in the baseline methods. We use a A100-8GB GPU machine for our experimentations. During the sampling process, we uniformly resize both the denoised image and the condition maps to dimensions of 512×512. In the context of color conditioning, we determine the optimal values for λ_1 and λ_2, utilized in the final loss term as outlined in Equation 11 of the main paper, to be 1 and 0.1 respectively. The reported results in our study are grounded in the application of these specified parameter values.

C. Limitations

In this Section, we briefly discuss a few limitations of TINTIN when seen in a conditional text-to-image generation setup. Firstly, TINTIN is slower than training-required methods during inference due to the presence of gradient computation of the energy function and iterative sampling strategy. Secondly, since the performance of TINTIN is highly dependent on the loss functions and the designated Conditioning Zone (CZ) for a specific condition, the extension of TINTIN to other conditions would require loss functions and the conditioning zone that can be very different for different conditions. Moreover, since we are building on top of existing text-to-image models, any potential fairness considerations for these base models will flow to our method as well.

Figure 2. We illustrate the ability of TINTIN in generating color palette conditioned images against trainable methods like T2I-Adapter[3] and ControlNet[2]. Our training-free approach is able to generate color balanced results as compared with other state-of-the-art methods that require training a model.

References

[1] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 1

[2] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer*

Figure 3. We illustrate the ability of TINTIN in generating edge map conditioned images against trainable methods like T2I-Adapter[3] and ControlNet[2]. Our training-free approach is able to generate diverse images following the structure of the reference edge map as compared with other state-of-the-art methods that require training a model.

Vision, 2023. 1, 2, 3

[3] 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 Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2024. 1, 2, 3

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz,

Patrick Esser, and Bjorn 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

[5] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1