[Supplementary Material]
Variational Distribution Learning for Unsupervised Text-to-Image Generation

Minsoo Kang[1*] Doyup Lee[3] Jiseob Kim[3] Saehoon Kim[3] Bohyung Han[1,2]
[1]ECE & [2]IPAI, Seoul National University [3]Kakao Brain
{kminsoo, bhhan}@snu.ac.kr {doyup.lee, jiseob.kim, shkim}@kakaobrain.com

A. Appendix

This document first provides the proof of Proposition 1. Then, we present additional results on the LN-COCO [4] dataset under the unsupervised T2I generation task based on StyleGAN2 [2]. We also evaluate the performance of the T2I results using a diffusion-based text-to-image generative model [6] to validate the generality of the proposed approach. Finally, we demonstrate additional qualitative results supplementing Figure 3 of the main paper, which visualizes the synthesized results on the MS-COCO [3] and Conceptual Captions 3M [8] datasets given by CLIP-GEN [9], LAFITE [10], and VDL based on StyleGAN2 under the unsupervised setting.

A.1. Proof of Proposition 1

Proposition 1. *Let $\hat{\mathbf{z}}_{txt}$ be the sample obtained by the proposed sampling strategy S_{VDL} based on $\mathbf{z}_{img}$. Then, the following inequality always holds for $G(\cdot;\phi^{\mathbf{z}_{txt}})$ with arbitrary values of its parameter $\phi^{\mathbf{z}_{txt}}$:*

$$\hat{\mathbf{z}}_{txt}^T \mathbf{z}_{img} \geq \sqrt{1-r^2},$$

where $r < 1$.

Proof. The inner product can be expressed as

$$\hat{\mathbf{z}}_{\text{txt}}^T \mathbf{z}_{\text{img}} = \mathbf{z}_{\text{img}}^T \frac{\mathbf{z}_{\text{img}} + r \cdot \mathbf{g}_{\text{img}}}{\|\mathbf{z}_{\text{img}} + r \cdot \mathbf{g}_{\text{img}}\|}, \tag{19}$$

where $\mathbf{g}_{\text{img}}$ is equivalent to Normalize$(G(\mathbf{z}_{\text{img}}))$. In addition, the denominator in (19) is given by

$$\begin{aligned}\|\mathbf{z}_{\text{img}} + r \cdot \mathbf{g}_{\text{img}}\| &= \sqrt{(\mathbf{z}_{\text{img}} + r \cdot \mathbf{g}_{\text{img}})^T (\mathbf{z}_{\text{img}} + r \cdot \mathbf{g}_{\text{img}})} \\ &= \sqrt{1 + 2r \cdot \mathbf{z}_{\text{img}}^T \mathbf{g}_{\text{img}} + r^2}. \end{aligned} \tag{20}$$

Based on the two equations, we have

$$\begin{aligned}\hat{\mathbf{z}}_{\text{txt}}^T \mathbf{z}_{\text{img}} &= \frac{\mathbf{z}_{\text{img}}^T (\mathbf{z}_{\text{img}} + r \cdot \mathbf{g}_{\text{img}})}{\sqrt{1 + 2r \cdot \mathbf{z}_{\text{img}}^T \mathbf{g}_{\text{img}} + r^2}} \\ &= \frac{1 + r \cdot \mathbf{z}_{\text{img}}^T \mathbf{g}_{\text{img}}}{\sqrt{1 + 2r \cdot \mathbf{z}_{\text{img}}^T \mathbf{g}_{\text{img}} + r^2}} \\ &= \frac{(1 + 2r \cdot \mathbf{z}_{\text{img}}^T \mathbf{g}_{\text{img}} + r^2) + (1 - r^2)}{2\sqrt{1 + 2r \cdot \mathbf{z}_{\text{img}}^T \mathbf{g}_{\text{img}} + r^2}} \\ &\geq \sqrt{1 - r^2}, \end{aligned}$$

where the last inequality is derived by using the inequality of arithmetic and geometric means. □

A.2. Experiments on LN-COCO

We compare the proposed method with CLIP-GEN [9] and LAFITE [10] on the LN-COCO [4] dataset using StyleGAN2 under the unsupervised setting. As presented in Table 5, our method outperforms existing approaches by large margins. Figure 7 depicts synthesized images given by VDL and CLIP-GEN [9], where the results of LAFITE are not provided since its pre-trained model is not publicly available.

A.3. Ablation on T2I models

We remark that our approach is model-agnostic; any type of conditional image generation network is applicable to our framework. To validate the effectiveness of the proposed method without carefully designing the T2I network, we replace StyleGAN2 [2] with a diffusion model, Latent Diffusion Model (LDM) [6], and perform experiments on the MS-COCO [3] and Conceptual Captions 3M [8] (CC3M) datasets under the unsupervised setting.

A.3.1 Implementation Details

For the second-stage training, we optimize LDM [6] for 150k and 300k iterations on the MS-COCO [3] and Con-

*This work was partly done during an internship at Kakao Brain.

ceptual Captions 3M [8] datasets using the Adam optimizer with a batch size of 64 and an initial learning rate of 6.4×10^{-5}. We set the resolution of the latent space to 64, where a pretrained Vector Quantized GAN [1] is selected as a latent perceptual compression network without an extra fine-tuning. Following the latent conditioning strategy used in [5], we inject CLIP features to the noisy predictions of the backbone network based on U-Net [7] inside the LDM framework [6]. Specifically, we replace the last group normalization layer in each residual block of the U-Net with an adaptive group normalization layer whose scale and shift parameters are computed by applying a single fully connected layer to the temporal positional embeddings. For conditioning given sentences, the outputs of the normalization layer are further multiplied with the projected CLIP features using a single fully connected layer.

A.3.2 Results unser Unsupervised Setting

We present quantitative results in Table 6 while the generated images on the MS-COCO [3] and Conceptual Captions 3M [8] datasets are provided in Figure 5 and 6, respectively. These results show that VDL archives superior performance when combined with the diffusion based LDM [6] for the T2I model, and the proposed method is agnostic to the types of the T2I network.

A.4. Additional Qualitative Results

Figure 7 and 8 visualize additional qualitative results from the proposed approach compared to existing methods. The results clearly show that VDL generates visually more faithful and realistic images considering the given text descriptions and the natural image distribution while the other two methods often fail to meet text conditions and/or generate natural images.

References

[1] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*, 2021. 2

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of Stylegan. In *CVPR*, 2020. 1, 3

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 1, 2, 3

[4] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting Vision and Language with Localized Narratives. In *ECCV*, 2020. 1, 3

[5] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *CVPR*, 2022. 2

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1, 2, 3

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 2

[8] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018. 1, 2, 3

[9] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP. *arXiv preprint arXiv:2203.00386*, 2022. 1, 3

[10] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards Language-Free Training for Text-to-Image Generation. In *CVPR*, 2022. 1, 3

Table 5. Results of unsupervised text-to-image generation on the LN-COCO [4] dataset using StyleGAN2 [2]. Methods with asterisks (*) report the results of our reproduction. A bold-faced number denotes the best performance in each column.

T2I Model	Dataset	Method	IS (↑)	FID (↓)	Sim_{txt} (↑)	Sim_{img} (↑)
StyleGAN2 [2]	LN-COCO [4]	CLIP-GEN* [9]	12.12	83.87	0.2750	–
		LAFITE [10]	18.49	38.95	0.0872	–
		VDL (Ours)	**21.55**	**31.33**	**0.6118**	**0.7025**

Figure 4. Qualitative results on the LN-COCO dataset using StyleGAN2. VDL generates visually higher-quality images than CLIP-GEN.

Table 6. Results of unsupervised text-to-image generation on the MS-COCO [3] and Conceptual Captions 3M [8] datasets using LDM [6].

T2I Model	Dataset	Method	IS (↑)	FID (↓)	Sim_{txt} (↑)	Sim_{img} (↑)
LDM [6]	MS-COCO [3]	CLIP-GEN* [9]	12.96	48.14	0.3042	-
		LAFITE* [10]	16.53	23.92	0.0965	-
		VDL (Ours)	**23.25**	**13.68**	**0.6104**	**0.7655**
	Conceptual Captions 3M [8]	CLIP-GEN* [9]	10.08	47.53	0.2896	-
		LAFITE* [10]	10.98	33.98	0.0912	-
		VDL (Ours)	**15.09**	**23.03**	**0.6237**	**0.7105**

Figure 5. Qualitative results on the MS-COCO dataset using LDM. VDL generates visually higher-quality images than LAFITE and CLIP-GEN.

Figure 6. Qualitative results on the Conceptual Captions 3M dataset using LDM. VDL generates visually higher-quality images than LAFITE and CLIP-GEN.

Figure 7. Additional qualitative results on the MS-COCO dataset using StyleGAN2. VDL generates visually higher-quality images than LAFITE and CLIP-GEN.

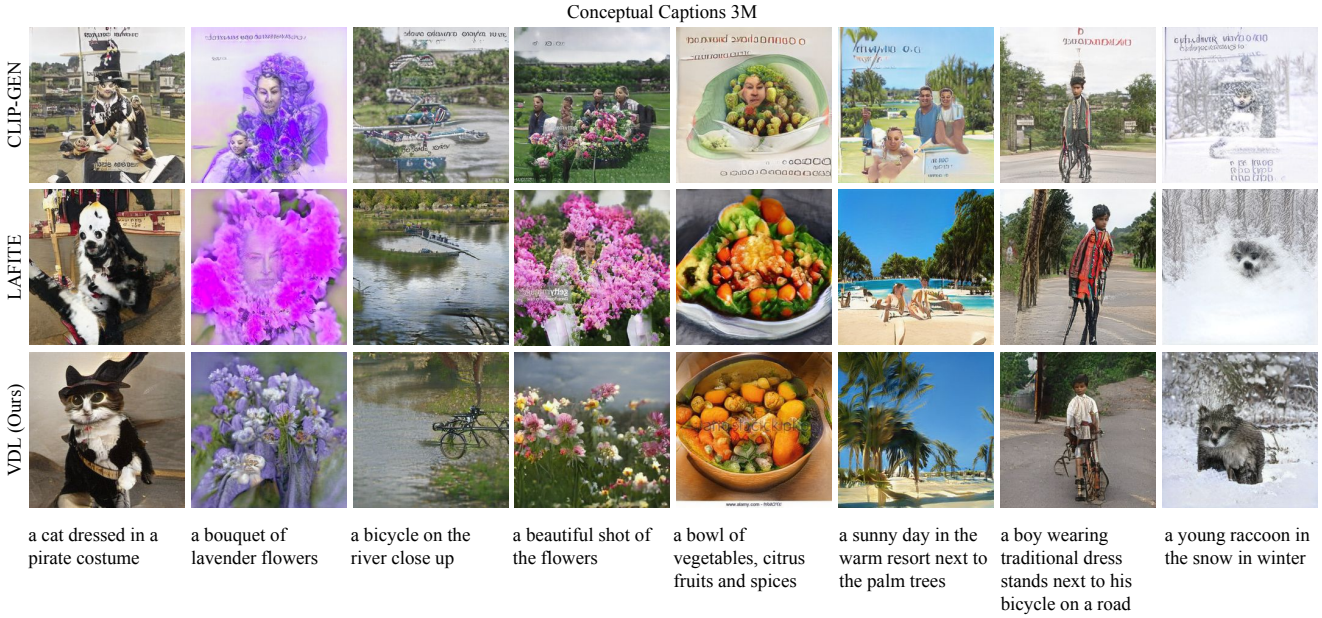

Figure 8. Additional qualitative results on the Conceptual Captions 3M dataset using StyleGAN2. VDL generates visually higher-quality images than LAFITE and CLIP-GEN.