Training-free Dense-Aligned Diffusion Guidance for Modular Conditional Image Synthesis

Supplementary Material

In this supplementary material, we provide additional information in the following three aspects:

A. discussion on synthesis with depth map and text;

B. quantitative analysis of synthesis with bounding box and text;

C. additional visualizations on various baselines.

A. Image Synthesis Results with Depth Map and Text

We evaluate performance based on descriptions and depth maps using ImageNet-R-TI2I dataset [9] as our benchmark. This dataset consists of 30 images across 10 object categories. Original images are converted into corresponding depth maps following the pipeline outlined in [11]. To quantify the accuracy of image synthesis, we employ Root Mean Squared Error (RMSE), which measures the pixel-wise discrepancy between the synthesized and original images. Since the given conditions specify a single foreground object, we utilize only the $\mathcal{L}_{cvg}$ component in the DGA module to optimize alignment and fidelity.

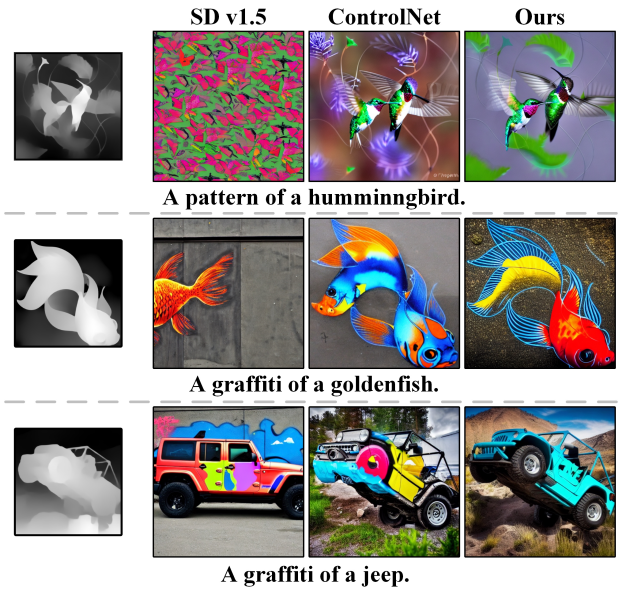

Figure 1. Visualization comparison between our approach (ControlNet [11] + DGA) and other baselines, under varying descriptions and depth maps. Our approach demonstrates a superior ability to adhere to the positional information within the depth map.

We use ControNet [11] as a baseline model to evaluate the effectiveness of our DGA module in aligning spatial configurations represented by depth maps. Tab. 1 summarized the RMSE score on ImageNet-R-TI2I dataset [9]. Our DGA module improves the performance over ControlNet by approximately 4.73%. This substantial gain is because our $\mathcal{L}_{cvg}$ component enhances the overlap between the object's layout and its depth map. Fig. 1 reveals a clear and consistent conclusion.

Table 1. Quantitative comparison under the textual description and depth map conditions. The data are derived from the ImageNet-R-TI2I dataset [9]. The evaluation metric used is RMSE. Since the conditions specify only one foreground object, we utilize the $\mathcal{L}_{cvg}$ component in the DGA module. The best results are **highlighted**.

Methods	Venue	RMSE ↓
Stable Diffusion [7]	CVPR2022	92.42
ControlNet [11]	CVPR2023	47.73
+ DGA	-	**45.47**

B. Additional Quantitative Results

Additional quantitative results of image synthesis with bounding box and text are shown in Tab. 2. Our DGA and DCA modules are also effective in this scenario.

Table 2. Quantitative comparison under the textual description and bounding box conditions provided by [6, 8]. The evaluation metric used is positional [1]. The baseline model of our approach is A&R [6]. The best results are **highlighted**.

Methods	Venue	Positional ↑
Stable Diffusion [7]	CVPR2022	12.50
Attend-and-Excite [3]	SIGGRAPH2023	20.50
DenseDiffusion [5]	ICCV2023	30.50
BoxDiff [10]	ICCV2023	32.50
MultiDiffusion [2]	ICML2023	36.00
Layout-guidance [4]	WACV2024	36.50
A&R [6]	CVPR2024	43.50
+ DGA	-	45.00
+ DGA + DCA	-	**47.50**

C. Additional Visualization Results

We conduct additional visualization comparison. We use Stable Diffusion v1.5 [7] as a baseline model to evalu-

ate the effectiveness of our DCA module in aligning concepts within descriptions. Visualization results are shown in Fig. 2. We use Dense Diffusion [5] as a baseline model to evaluate the effectiveness of our DGA module in aligning spatial configurations represented by segmentation masks; and our DCA module in aligning geometric concepts within descriptions. Visualization results are shown in Fig. 3.

As shown in Fig. 2, our approach accurately generates multiple objects, *e.g.*, two bears (row 1), two people with a parachute (row 2), and a couple with a dog (row 4), demonstrating its superior compositional reasoning capability. Additionally, our model effectively synthesizes small yet important elements, such as the penguin napkin holder (row 3), highlighting its enhanced ability to preserve fine details. Furthermore, the objects in our results exhibit more natural interactions with their environment, *e.g.*, the polar bears engaging with water (row 1) and the parachute partially resting on the sand (row 2), indicating improved spatial and contextual alignment.

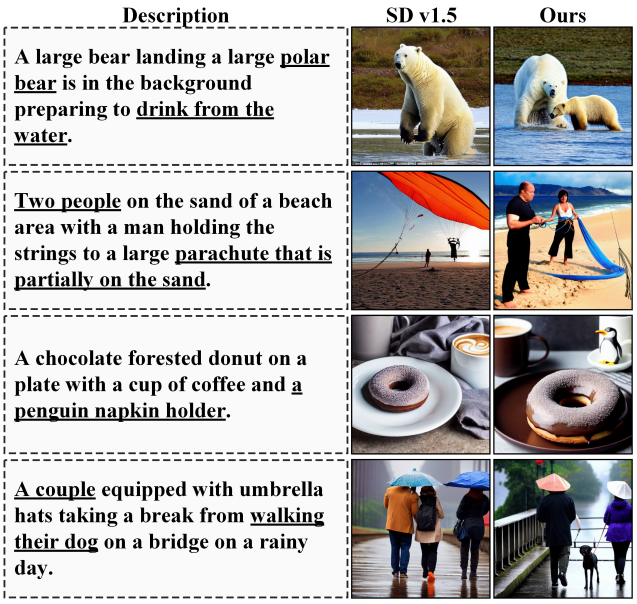

Figure 2. Visualization comparison between our approach (Stable Diffusion v1.5 [7] + DCA) and its baseline under varying descriptions. Our approach exhibits superior compositional reasoning, detail preservation, and contextual alignment.

As shown in Fig. 3, our approach ensures that multiple objects are correctly positioned relative to each other, *e.g.*, the surfboard is properly placed on the car roof (row 1), and the dog is accurately seated on the bench (row 2), demonstrating its superior object placement capability. Additionally, our generated objects maintain coherent geometry and proportions, *e.g.*, the cat and laptop are well-defined and spatially distinguishable (row 3), highlighting the model's ability to preserve structural consistency. Furthermore, our method effectively respects depth relationships, *e.g.*, the couch and TV exhibit a realistic depth ordering (row 4), ensuring proper scene composition.

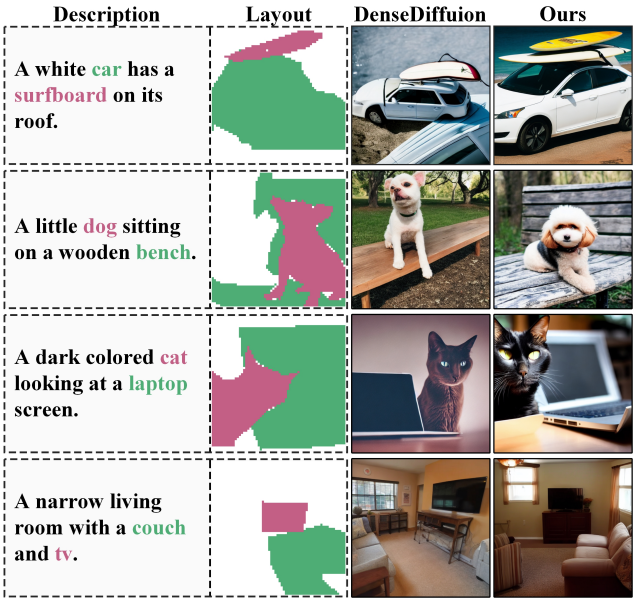

Figure 3. Visualization comparison between our approach (Dense Diffusion [5] + DGA+ DCA) and its baseline, under varying descriptions and segmentation masks.. Our approach can satisfy both conditions.

References

[1] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20041–20053, 2023. 1

[2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning (ICML)*, pages 1737–1752. PMLR, 2023. 1

[3] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1

[4] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5343–5353, 2024. 1

[5] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7701–7711, 2023. 1, 2

[6] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7932–7942, 2024. 1

[7] 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 (CVPR)*, pages 10684–10695, 2022. 1, 2

[8] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 1

[9] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 1

[10] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7452–7461, 2023. 1

[11] 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 (CVPR)*, pages 3836–3847, 2023. 1