Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors Supplementary

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman

Meta AI Research

A Additional implementation details

The high-level architecture is depicted in Fig. 1.

VQ-SEG. VQ-SEG is trained for $600k$ iterations, with a batch size of 48, dictionary size of 1024. The number of segmentation categories per-group are $m_p = 133$ for the panoptic segmentation, $m_h = 20$ for the human parsing, and $m_f = 5$ for the face parsing. The per-category weight function follows the notation:

$$\alpha_{cat} = \begin{cases} 20, & \text{if cat} \in [154, ..., 158] \\ 1, & \text{otherwise,} \end{cases} \tag{1}$$

where cat $\in [154, ..., 158]$ are the face-parts categories eyebrows, eyes, nose, outer-mouth, and inner-mouth.

VQ-IMG. VQ-IMG$_{256}$ and VQ-IMG$_{512}$ are trained for $800k$ and $940k$ iterations respectively, with a batch size of 192 and 128, a channel multiplier of $[1, 1, 2, 4]$ and $[1, 1, 2, 4, 4]$, while both are trained with a dictionary size of 8192.

The per-layer normalizing hyperparameter for the face-aware loss is $\alpha_f^l = [\alpha_{f1}, \alpha_{f2} \times 0.01, \alpha_{f2} \times 0.1, \alpha_{f2} \times 0.2, \alpha_{f2} \times 0.02]$ corresponding to the last layer of each block of size $1 \times 1, 7 \times 7, 28 \times 28, 56 \times 56, 128 \times 128$, where $\alpha_{f1} = 0.1$ and $\alpha_{f2} = 0.25$. We experimented with two settings, the first where $\alpha_{f1} = \alpha_{f2} = 1.0$, and the second, which was used to train the final models, where $\alpha_{f1} = 0.1, \alpha_{f2} = 0.25$. The remaining face-loss values were taken from the work of [2]. The per-layer normalizing hyperparameter for the object-aware loss, α_o^l were taken from the work of [1], based on LPIPS [4].

Scene-based transformer. The 512×512 and 256×256 models both share all implementation details, excluding the VQ-IMG used for token encoding and decoding, and the object-aware loss that was applied to the 512×512 model only. Both transformers share the architecture of 48 layers, 48 attention heads, and an embedding dimension of 2560. The models were trained for a total of $170k$ iterations, with a batch size of 1024, Adam [3] optimizer, with a starting

learning-rate of 4.5×10^{-4} for the first $40k$ iterations, transitioning to 1.5×10^{-4} for the remainder, $\beta_1 = 0.9$,$\beta_2 = 0.96$, weight-decay of 4.5×10^{-4}, and a loss ratio of 7/1 between the image and text tokens. For classifier-free guidance, we fine-tune the transformer, while replacing the text tokens with padding tokens in the last $30k$ iterations, with a probability of $p_{CF} = 0.2$. At inference-time we set the guidance scale to $\alpha_c = 5$, though we found that $\alpha_c = 3$ works as well.

At each inference step, the next token is sampled by (i) selecting half the logits with the highest probabilities, (ii) applying a softmax operation over the selected logits, and (iii) sampling a single logit from a multinomial probability distribution.

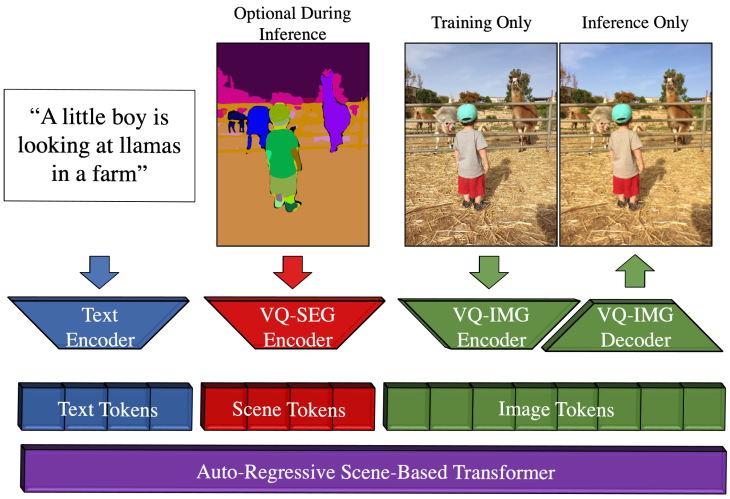

Fig. 1. The scene-based method high-level architecture. Given an input text and optional scene layout, a corresponding image is generated. The transformer generates the relevant tokens, encoded and decoded by the corresponding networks.

B Additional samples

Additional samples generated from challenging text inputs are provided in Figs. 2-3, while samples generated from text and scene inputs are provided in Figs. 4-7. The different text colors emphasize the large number of different objects/scenarios being attended. As there are no 'octopus' or 'dinosaur' categories, we use instead the 'cat' and 'giraffe' categories respectively. We did not attempt to use other classes in this case. However, we found that generally there are no "one-to-one" mappings between absent and existing categories, hence several categories may work for an absent category.

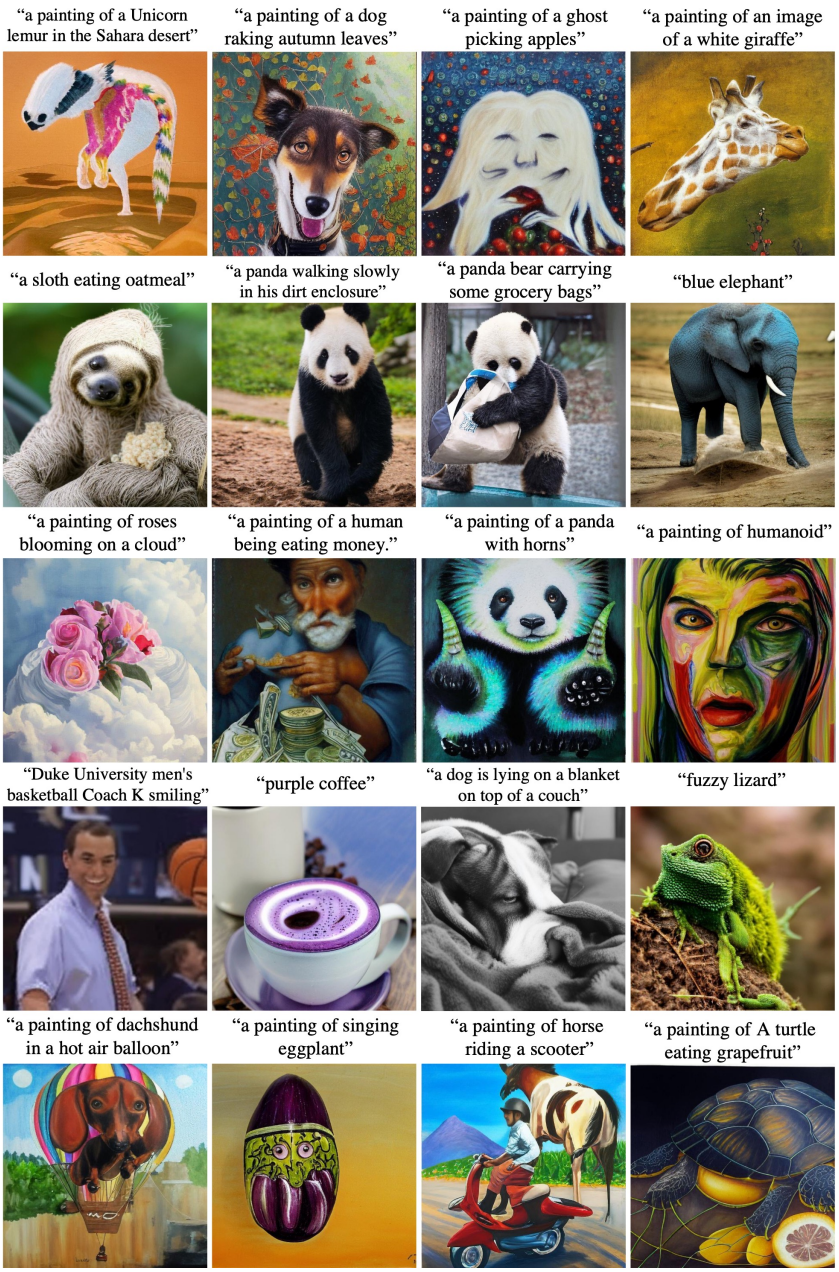

Fig. 2. Additional samples generated from challenging text inputs.

Fig. 3. Additional samples generated from challenging text inputs.

Fig. 4. Additional samples generated (b) from text and segmentation inputs (a).

C Additional comparisons and ablation studies

A comparison of challenging generations (faces) is shown in Fig. 8. This challenging comparison emphasizes the necessity of the face-aware component in the VQ-IMG network.
To further establish the necessity of model large-scaling, we provide an ablation study of two smaller models (Small-0.4B, Medium-1.4B) in Tab. 1.

Model	FID↓	Text↑	Quality↑
Obj_{512}	8.70	-	-
+S / M	19.08 / 15.49	34.4% / 39.5%	30.5% / 42.9%

Table 1. An ablation study (FID and human preference) of the small (S, 0.4B parameters) and medium (M, 1.4B) models. Human preference (text-alignment and image-quality) of S and M are compared with Obj_{512} (4B model).

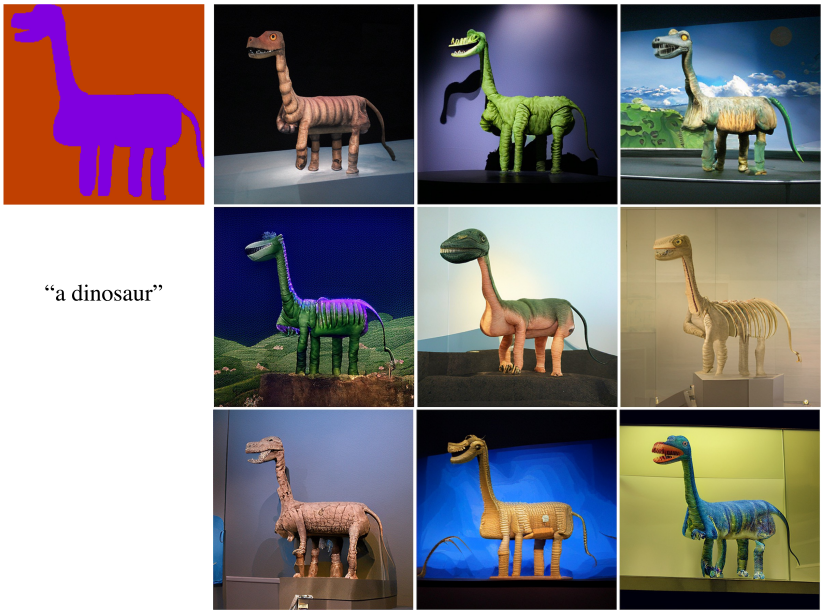

Fig. 5. Additional samples generated (b) from text and segmentation inputs (a).

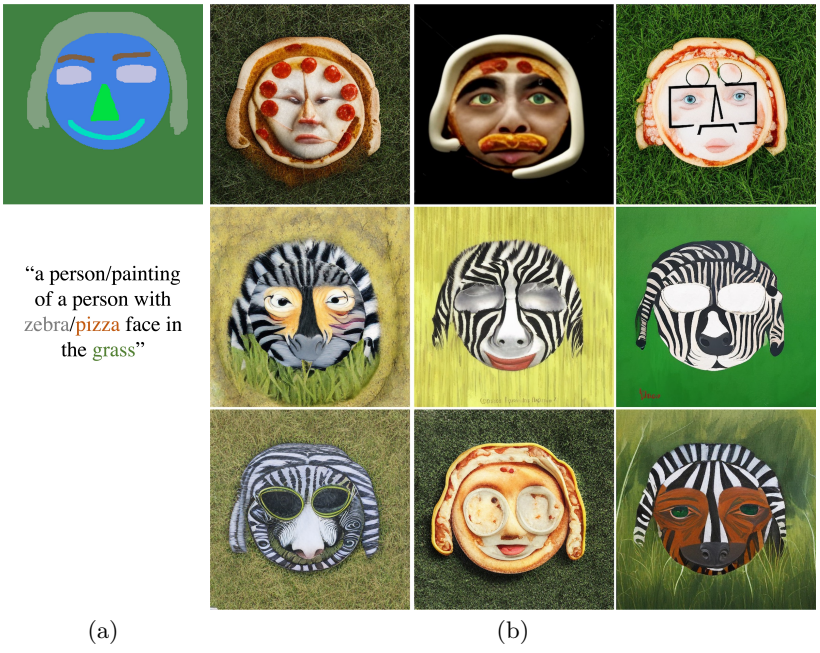

Fig. 6. Additional samples generated (b) from text and segmentation inputs (a).

Fig. 7. Additional samples generated (b) from text and segmentation inputs (a).

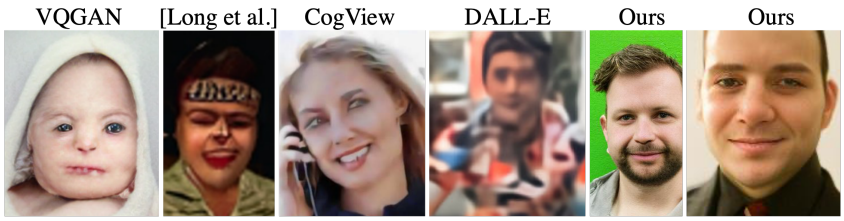

Fig. 8. A qualitative comparison emphasizing the advantage of our method in challenging generations (where available). [5] is denoted as Long et al.

References

1. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021)
2. Gafni, O., Wolf, L., Taigman, Y.: Live face de-identification in video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9378–9387 (2019)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
5. Zhao, L., Zhang, Z., Chen, T., Metaxas, D., Zhang, H.: Improved transformer for high-resolution gans. Advances in Neural Information Processing Systems **34**, 18367–18380 (2021)