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

Oran Gafni Adam Polyak Oron Ashual Shelly Sheynin Devi Parikh Yaniv Taigman
Meta AI Research
{oran,adampolyak,oron,shellysheynin,dparikh,yaniv}@fb.com

(a) (b)

Figure 1. Make-A-Scene: Samples of generated images from text inputs (a), and a text and scene input (b). Our method is able to both generate the scene (a, bottom left) and image, or generate the image from text and a simple sketch input (b, center).

Abstract

Recent text-to-image generation methods provide a simple yet exciting conversion capability between text and image domains. While these methods have incrementally improved the generated image fidelity and text relevancy, several pivotal gaps remain unanswered, limiting applicability and quality. We propose a novel text-to-image method that addresses these gaps by (i) enabling a simple control mechanism complementary to text in the form of a scene, (ii) introducing elements that substantially improve the tokenization process by employing domain-specific knowledge over key image regions (faces and salient objects), and (iii) adapting classifier-free guidance for the transformer use case. Our model achieves state-of-the-art FID and human evaluation results, unlocking the ability to generate high fidelity images in a resolution of 512×512 *pixels, significantly improving visual quality. Through scene controllability, we introduce several new capabilities:* ***(i)*** *Scene editing,* ***(ii)*** *text editing with anchor scenes,* ***(iii)*** *overcoming out-of-distribution text prompts, and* ***(iv)*** *story illustration generation, as demonstrated in the story we wrote.*

1. Introduction

"A poet would be overcome by sleep and hunger before being able to describe with words what a painter is able to depict in an instant."

Similar to this quote by Leonardo da Vinci [27], equivalents of the expression "A picture is worth a thousand words" have been iterated in different languages and eras [14, 1, 25], alluding to the heightened expressiveness of images over text, from the human perspective. There is no surprise then, that the task of text-to-image generation has been gaining increased attention with the recent success of text-to-image modeling via large-scale models and datasets. This new capability of effortlessly bridging between the text and image domains enables new forms of creativity to be accessible to the general public.

While current methods provide a simple yet exciting conversion between the text and image domains, they still lack several pivotal aspects:

(i) Controllability. The sole input accepted by the majority of models is text, confining any output to be controlled by a text description only. While certain perspectives

can be controlled with text, such as style or color, others such as structure, form, or arrangement can only be loosely described at best [46]. This lack of control conveys a notion of randomness and weak user-influence on the image content and context [34]. Controlling elements additional to text have been suggested by [69], yet their use is confined to restricted datasets such as fashion items or faces. An earlier work by [23] suggests coarse control in the form of bounding boxes resulting in low resolution images.

(ii) Human perception. While images are generated to match human perception and attention, the generation process does not include any relevant prior knowledge, resulting in little correlation between generation and human attention. A clear example of this gap can be observed in person and face generation, where a dissonance is present between the importance of face pixels from the human perspective and the loss applied over the whole image [28, 66]. This gap is relevant to animals and salient objects as well.

(iii) Quality and resolution. Although quality has gradually improved between consecutive methods, the previous state-of-the-art methods are still limited to an output image resolution of 256×256 pixels [45, 41]. Alternative approaches propose a super-resolution network which results in less favorable visual and quantitative results [12]. Quality and resolution are strongly linked, as scaling up to a resolution of 512×512 requires a substantially higher quality with fewer artifacts than 256×256.

In this work, we introduce a novel method that successfully tackles these pivotal gaps, while attaining state-of-the-art results in the task of text-to-image generation. Our method provides a new type of control complementary to text, enabling new-generation capabilities while improving structural consistency and quality. Furthermore, we propose explicit losses correlated with human preferences, significantly improving image quality, breaking the common resolution barrier, and thus producing results in a resolution of 512×512 pixels.

Our method is comprised of an autoregressive transformer, where in addition to the conventional use of text and image tokens, we introduce implicit conditioning over optionally controlled scene tokens, derived from segmentation maps. During inference, the segmentation tokens are either generated independently by the transformer or extracted from an input image, providing freedom to impel additional constraints over the generated image. Contrary to the common use of segmentation for explicit conditioning as employed in many GAN-based methods [24, 62, 42], our segmentation tokens provide implicit conditioning in the sense that the generated image and image tokens are not constrained to use the segmentation information, as there is no loss tying them together. In practice, this contributes to the variety of samples generated by the model, producing diverse results constrained to the input segmentations.

We demonstrate the new capabilities this method provides in addition to controllability, such as (i) complex scene generation (Fig. 1), (ii) out-of-distribution generation (Fig. 3), (iii) scene editing (Fig. 4), and (iv) text editing with anchored scenes (Fig. 5). We additionally provide an example of harnessing controllability to assist with the creative process of storytelling in this video.

While most approaches rely on losses agnostic to human perception, this approach differs in that respect. We use two modified Vector-Quantized Variational Autoencoders (VQ-VAE) to encode and decode the image and scene tokens with explicit losses targeted at specific image regions correlated with human perception and attention, such as faces and salient objects. The losses contribute to the generation process by emphasizing the specific regions of interest and integrating domain-specific perceptual knowledge in the form of network feature-matching.

While some methods rely on image re-ranking for post-generation image filtering (utilizing CLIP [44] for instance), we extend the use of classifier-free guidance suggested for diffusion models [53, 20] by [22, 41] to transformers, eliminating the need for post-generation filtering, thus producing faster and higher quality generation results, better adhering to input text prompts.

An extensive set of experiments is provided to establish the visual and numerical validity of our contributions.

2. Related Work

2.1. Image generation

Recent advancements in deep generative models have enabled algorithms to generate high-quality and natural-looking images. Generative Adversarial Networks (GANs) [17] facilitate the generation of high fidelity images [29, 3, 30, 56] in multiple domains by simultaneously training a generator network G and a discriminator network D, where G is trained to fool D, while D is trained to judge if a given image is real or fake. Concurrently to GANs, Variational Autoencoders (VAEs) [32, 57] have introduced a likelihood-based approach to image generation. Other likelihood-based models include autoregressive models [58, 43, 13, 8] and diffusion models [11, 21, 20]. While the former model image pixels as a sequence with autoregressive dependency between each pixel, the latter synthesizes images via a gradual denoising process. Specifically, sampling starts with a noisy image which is iteratively denoised until all denoising steps are performed. Applying both methods directly to the image pixel-space can be challenging. Consequently, recent approaches either compress the image to a discrete representation [13, 59] via Vector Quantized (VQ) VAEs [59], or down-sample the image resolution [11, 21]. Our method is based on autoregressive modeling of discrete image representation.

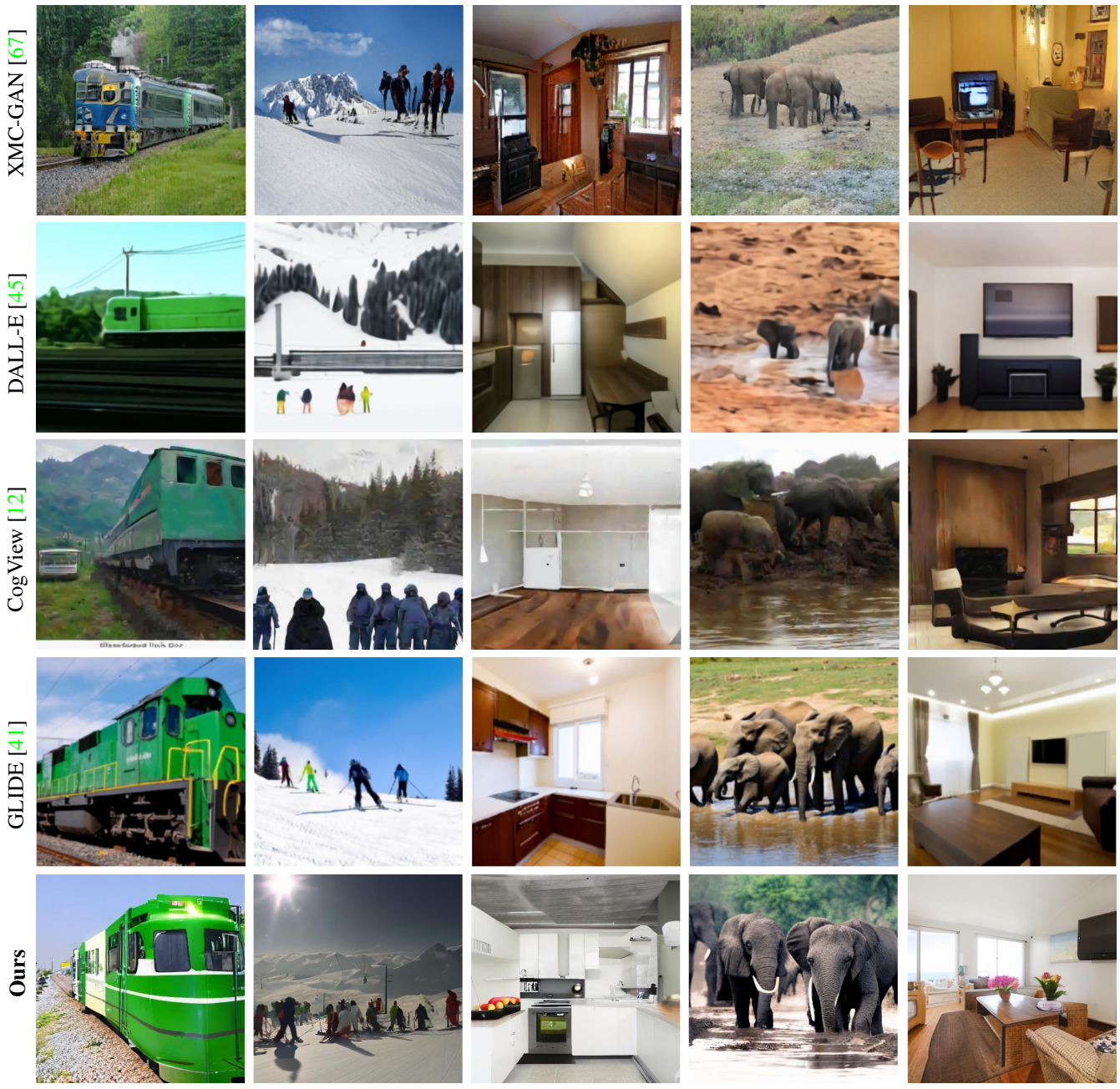


Figure 2. Qualitative comparison with previous work. The text and generated images for [67, 45, 41] were taken from [41]. For CogView [12] we use the released 512×512 model weights, applying self-reranking of 60 for post-generation selection.

2.2. Image tokenization

Image generation models based on discrete representation [59, 45, 47, 12, 13] follow a two-stage training scheme. First, an image tokenizer is trained to extract a discrete image representation. In the second stage, a generative model generates the image in the discrete latent space. Inspired by Vector Quantization (VQ) techniques, VQ-VAE [59] learns to extract a discrete latent representation by performing online clustering. VQ-VAE-2 [47] presented a hierarchical ar-

chitecture composed of VQ-VAE models operating at multiple scales, enabling faster generation compared with pixel space generation. The DALL-E [45] text-to-image model used dVAE, which uses gumbel-softmax [26, 39], relaxing the VQ-VAE's online clustering. Recently, VQGAN [13] added adversarial and perceptual losses [68] on top of the VQ-VAE reconstruction task, producing reconstructed images with higher quality. In our work, we modify the VQ-GAN framework by adding perceptual losses to specific image regions, such as faces and salient objects, which further improve the fidelity of the generated images.

2.3. Image-to-image generation

Generating images from segmentation maps or scenes can be viewed as a conditional image synthesis task [71, 38, 24, 61, 62, 42]. Specifically, this form of image synthesis permits more controllability over the desired output. CycleGAN [71] trained a mapping function from one domain to the other. UNIT [38] projected two different domains into a shared latent space and used a per-domain decoder to re-synthesize images in the desired domain. Both methods do not require supervision between domains. pix2pix [24] utilized conditional GANs together with a supervised reconstruction loss. pix2pixHD [62] improved the latter by increasing output image resolution thanks to improved network architecture. SPADE [42] introduced a spatially-adaptive normalization layer which elevated information lost in normalization layers. [15] introduced face-refinement to SPADE through a pre-trained face-embedding network inspired by face-generation methods [16]. Unlike the aforementioned, our work conditions jointly on text and segmentation, enabling bi-domain controllability.

2.4. Text-to-image generation

Text-to-image generation [64, 72, 54, 65, 67, 45, 12, 41, 70] focuses on generating images from standalone text descriptions. Preliminary text-to-image methods conditioned RNN-based DRAW [18] on text [40]. Text-conditioned GANs provided additional improvement [48]. AttnGAN [64] introduced an attention component, allowing the generator network to attend to relevant words in the text. DM-GAN [72] introduced a dynamic memory component, while DF-GAN [54] employed a fusion block, fusing text information into image features. Contrastive learning further improved the results of DM-GAN [65], while XMC-GAN [67] used contrastive learning to maximize the mutual information between image and text.

DALL-E [45] and CogView [12] trained an autoregressive transformer [60] on text and image tokens, demonstrating convincing zero-shot capabilities on the MS-COCO dataset. GLIDE [41] used diffusion models conditioned on images. Inspired by the high-quality unconditional images generation model, GLIDE employed guided inference with and without a classifier network to generate high-fidelity images. LAFITE [70] employed a pre-trained CLIP [44] model to project text and images to the same latent space, training text-to-image models without text data. Similarly to DALL-E and CogView, we train an autoregressive transformer model on text and image tokens. Our main contributions are introducing additional controlling elements in the form of a scene, improve the tokenization process, and adapt classifier-free guidance to transformers.

3. Method

Our model generates an image given a text input and an optional scene layout (segmentation map). As demonstrated in our experiments, by conditioning over the scene layout, our method provides a new form of implicit controllability, improves structural consistency and quality, and adheres to human preference (as assessed by our human evaluation study). In addition to our scene-based approach, we extended our aspiration of improving the general and perceived quality with a better representation of the token space. We introduce several modifications to the tokenization process, emphasizing awareness of aspects with increased importance in the human perspective, such as faces and salient objects. To refrain from post-generation filtering and further improve the generation quality and text alignment, we employ classifier-free guidance.

We follow next with a detailed overview of the proposed method, comprised of (i) scene representation and tokenization, (ii) attending human preference in the token space with explicit losses, (iii) the scene-based transformer, and (iv) transformer classifier-free guidance. Aspects commonly used prior to this method are not extensively detailed below, whereas specific settings for all elements can be found in the appendix.

3.1. Scene representation and tokenization

The scene is composed of a union of three complementary semantic segmentation groups - panoptic, human, and face. By combining the three extracted semantic segmentation groups, the network learns to both generate the semantic layout and condition on it while generating the final image. The semantic layout provides additional global context in an implicit form that correlates with human preference, as the choice of categories within the scene groups, and the choice of the groups themselves are a prior to human preference and awareness. We consider this form of conditioning to be implicit, as the network may disregard any scene information, and generate the image conditioned solely on text. Our experiments indicate that both the text and scene firmly control the image.

In order to create the scene token space, we employ VQ-SEG: a modified VQ-VAE for semantic segmentation, building on the VQ-VAE suggested for semantic segmen-

Figure 3. Overcoming out-of-distribution text prompts with scene control. By introducing simple scene sketches (bottom right) as additional inputs, our method is able to overcome unusual objects and scenarios presented as failure cases in previous methods.

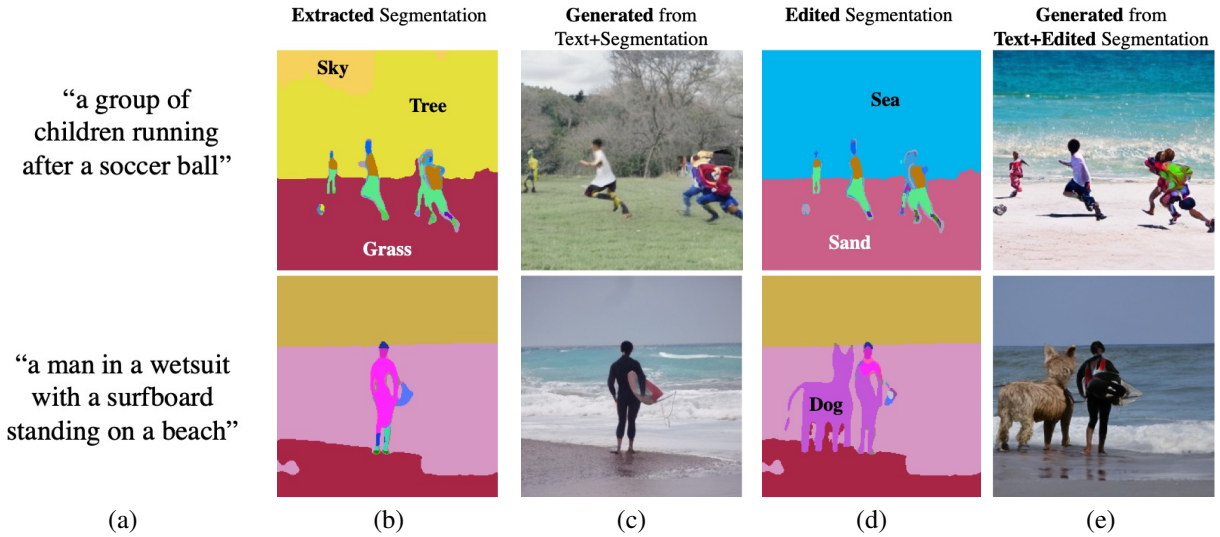


Figure 4. Generating images through edited scenes. For an input text (a) and the segmentations extracted from an input image (b), we can re-generate the image (c) or edit the segmentations (d) by replacing classes (top) or adding classes (bottom), generating images with new context or content (e).

tation in [13]. In our implementation the inputs and outputs of VQ-SEG are m channels, representing the number of classes for all semantic segmentation groups $m = m_p + m_h + m_f + 1$, where m_p, m_h, m_f are the number of categories for the panoptic segmentation [63], human segmentation [35], and face segmentation extracted with [5] respectively. The additional channel is a map of the edges separating the different classes and instances. The edge channel provides both separations for adjacent instances of the same class, and emphasis on scarce classes with high importance, as edges (perimeter) are less biased towards larger categories than pixels (area).

3.2. Adhering to human emphasis in the token space

We observe an inherent upper-bound on image quality when generating images with the transformer, stem-

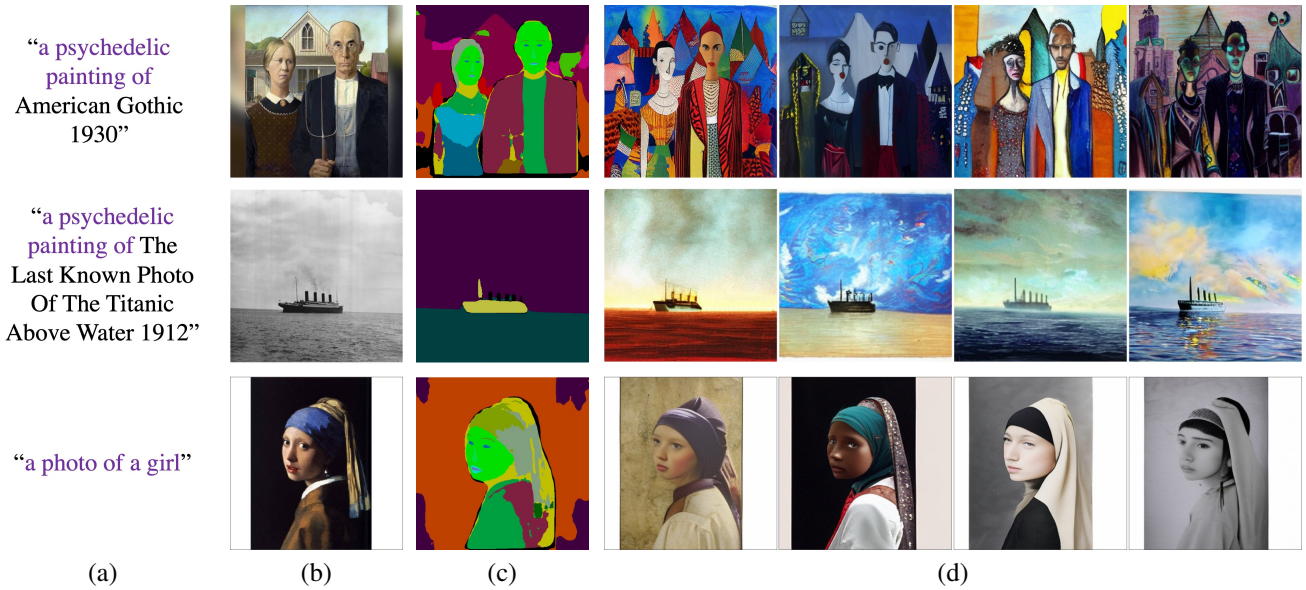


Figure 5. Generating new image interpretations through text editing and anchor scenes. For an input text (a) and image (b), we first extract the semantic segmentation (c), we can then re-generate new images (d) given the input segmentation and edited text. Purple denotes text added or replacing the original text.

ming from the tokenization reconstruction method. In other words, quality limitations of the VQ image reconstruction method inherently transfer to quality limitations on images generated by the transformer. To that end, we introduce several modifications to both the segmentation and image reconstruction methods. These modifications are losses in the form of emphasis (specific region awareness) and perceptual knowledge (feature-matching over task-specific pre-trained networks).

3.3. Face-aware vector quantization

While using a scene as an additional form of conditioning provides an implicit prior for human preference, we institute explicit emphasis in the form of additional losses, explicitly targeted at specific image regions.

We employ a feature-matching loss over the activations of a pre-trained face-embedding network, introducing "awareness" of face regions and additional perceptual information, motivating high-quality face reconstruction.

Before training the face-aware VQ (denoted as VQ-IMG), faces are located using the semantic segmentation information extracted for VQ-SEG. The face locations are then used during the face-aware VQ training stage, running up to k_f faces per image from the ground-truth and reconstructed images through the face-embedding network. The face loss can then be formulated as following:

$$\mathcal{L}_{\text{Face}} = \sum_k \sum_l \alpha_f^l \| \text{FE}^l(\hat{c}_f^k) - \text{FE}^l(c_f^k) \|, \qquad (1)$$

where the index l is used to denote the size of the spatial activation at specific layers of the face embedding network FE [6], while the summation runs over the last layers of each block of size 112×112, 56×56, 28×28, 7×7, 1×1 (1×1 being the size of the top most block), $\hat{c}_f^k$ and c_f^k are respectively the reconstructed and ground-truth face crops k out of k_f faces in an image, α_f^l is a per-layer normalizing hyperparameter, and $\mathcal{L}_{\text{Face}}$ is the face loss added to the VQGAN losses defined by [13].

3.4. Face emphasis in the scene space

While training the VQ-SEG network, we observe a frequent reduction of the semantic segmentations representing the face parts (such as the eyes, nose, lips, eyebrows) in the reconstructed scene. This effect is not surprising due to the relatively small number of pixels that each face part accounts for in the scene space. A straightforward solution would be to employ a loss more suitable for class imbalance, such as focal loss [36]. However, we do not aspire to increase the importance of classes that are both scarce, and of less importance, such as fruit or a tooth-brush. Instead, we (1) employ a weighted binary cross-entropy face loss over the segmentation face parts classes, emphasizing higher importance for face parts, and (2) include the face parts edges as part of the semantic segmentation edge map mentioned above. The weighted binary cross-entropy loss can then be formulated as following:

$$\mathcal{L}_{\text{WBCE}} = \alpha_{cat} \, \text{BCE}(s, \hat{s}), \qquad (2)$$

where s and $\hat{s}$ are the input and reconstructed segmentation maps respectively, α_{cat} is a per-category weight function, BCE is a binary cross-entropy loss, and $\mathcal{L}_{\text{WBCE}}$ is the

weighted binary cross-entropy loss added to the conditional VQ-VAE losses defined by [13].

3.5. Object-aware vector quantization

We generalized and extend the face-aware VQ method to increase awareness and perceptual knowledge of objects defined as "things" in the panoptic segmentation categories. Rather than a specialized face-embedding network, we employ a pre-trained VGG [52] network trained on ImageNet [33], and introduce a feature-matching loss representing the perceptual differences between the object crops of the reconstructed and ground-truth images. By running the feature-matching over image crops, we are able to increase the output image resolution from 256×256 by simply adding to VQ-IMG an additional down-sample and up-sample layer to the encoder and decoder respectively. Similarly to Eq. 1, the loss can be formulated as:

$$\mathcal{L}_{\text{Obj}} = \sum_k \sum_l \alpha_o^l \| \text{VGG}^l(\hat{c}_o^k) - \text{VGG}^l(c_o^k) \|, \quad (3)$$

where $\hat{c}_o^k$ and c_o^k are the reconstructed and input object crops respectively, VGG^l are the activations of the $l-th$ layer from the pre-trained VGG network, α_o^l is a per-layer normalizing hyperparameter, and $\mathcal{L}_{\text{Obj}}$ is the object-aware loss added to the VQ-IMG losses defined in Eq. 1.

3.6. Scene-based transformer

The method relies on an autoregressive transformer with three independent consecutive token spaces: text, scene, and image, as depicted in Fig 6. The token sequence is comprised of n_x text tokens encoded by a BPE [50] encoder, followed by n_y scene tokens encoded by VQ-SEG, and n_z image tokens encoded or decoded by VQ-IMG.

Prior to training the scene-based transformer, each encoded token sequence corresponding to a [text, scene, image] triplet is extracted using the corresponding encoder, producing a sequence that consists of:

$$t_x, t_y, t_z = \text{BPE}(i_x), \text{VQ-SEG}(i_y), \text{VQ-IMG}(i_z),$$
$$t = [t_x, t_y, t_z],$$

where i_x, i_y, i_z are the input text, scene and image respectively, $i_x \in \mathbb{N}^{d_x}$, d_x is the length of the input text sequence, $i_y \in \mathbb{R}^{h_y \times w_y \times m}$, $i_z \in \mathbb{R}^{h_z \times w_z \times 3}$, h_y, w_y, h_z, w_z are the height and width dimensions of the scene and image inputs respectively, BPE is the Byte Pair Encoding encoder, t_x, t_y, t_z are the text, scene and image input tokens respectively, and t is the complete token sequence.

3.7. Transformer classifier-free guidance

Inspired by the high-fidelity of unconditional image generation models, we employ classifier-free guidance [9, 22, 44]. Classifier-free guidance is the process of guiding an unconditional sample in the direction of a conditional sample. To support unconditional sampling we fine-tune the transformer while randomly replacing the text prompt with padding tokens with a probability of p_{CF}. During inference, we generate two parallel token streams: a conditional token stream conditioned on text, and an unconditional token stream conditioned on an empty text stream initialized with padding tokens. For transformers, we apply classifier-free guidance on logit scores:

$$\text{logits}_{cond} = T(t_y, t_z | t_x),$$
$$\text{logits}_{uncond} = T(t_y, t_z | \emptyset),$$
$$\text{logits}_{cf} = \text{logits}_{uncond} + \alpha_c \cdot (\text{logits}_{cond} - \text{logits}_{uncond}),$$

where $\emptyset$ is the empty text stream, $logits_{cond}$ are logit scores outputted by the conditioned token stream, $logits_{uncond}$ are logit scores outputted by the unconditioned token stream, α_c is the guidance scale, $logits_{cf}$ is the guided logit scores used to sample the next scene or image token, T is an autoregressive transformer based the GPT-3 [4] architecture. Note that since we use an autoregressive transformer, we use $logits_{cf}$ to sample once and feed the same token (image or scene) to the conditional and unconditional stream.

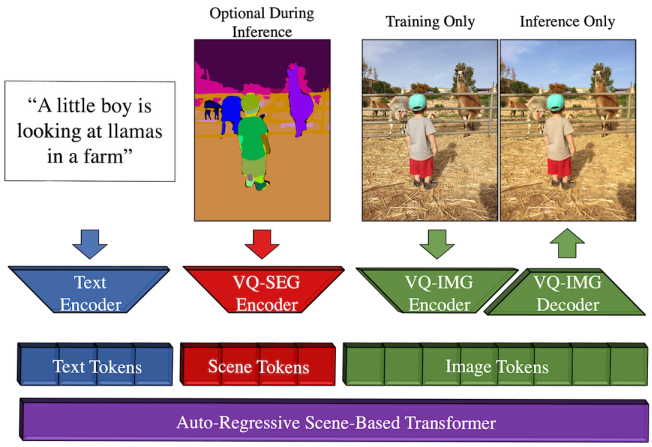


Figure 6. The scene-based method high-level architecture. Given an input text and optional scene layout, a corresponding image is generated. The transformer generates the relevant tokens, encoded and decoded by the corresponding networks.

4. Experiments

Our model achieves state-of-the-art results in human-based and numerical metric comparisons. Samples supporting the qualitative advantage are provided in Fig. 2. Additionally, we demonstrate new creative capabilities possible with this method's new form of controllability. Finally, to better assess the effect of each contribution, an ablation study is provided.

Experiments were performed with a 4 billion parameter transformer, generating a sequence of 256 text tokens, 256 scene tokens, and 1024 image tokens, that are then decoded into an image with a resolution of 256×256 or 512×512 pixels (depending on the model of choice).

4.1. Datasets

The scene-based transformer is trained on a union of CC12m [7], CC [51], and subsets of YFCC100m [55] and Redcaps [10], amounting to 35m text-image pairs. MS-COCO [37] is used unless otherwise specified. VQ-SEG and VQ-IMG are trained on CC12m, CC, and MS-COCO.

4.2. Metrics

The goal of text-to-image generation is to generate high-quality and text-aligned images from a human perspective. Different metrics have been suggested to mimic the human perspective, where some are considered more reliable than others. We consider human evaluation the highest authority when evaluating image quality and text-alignment, and rely on FID [19] to increase evaluation confidence and handle cases where human evaluation is not applicable. We do not use IS [49] as it has been noted to be insufficient for model evaluation [2].

4.3. Comparison with previous work

The task of text-to-image generation does not contain absolute ground-truths, as a specific text description could apply to multiple images and vice versa. This constrains evaluation metrics to evaluate distributions of images, rather than specific images, thus we employ FID [19] as our secondary metric.

4.4. Baselines

We compare our results with several state-of-the-art methods using the FID metric and human evaluators (AMT) when possible. **DALL-E** [45] provides strong zero-shot capabilities, similarly employing an autoregressive transformer with VQ-VAE tokenization. We train a re-implementation of DALL-E with 4B parameters to enable human evaluation and fairly compare both methods employing an identical VQ method (VQGAN). **GLIDE** [41] demonstrates vastly improved results over DALL-E, adopting a diffusion-based [53] approach with classifier-free guidance [22]. We additionally provide an FID comparison with **CogView** [12], **LAFITE** [70], **XMC-GAN** [67], **DM-GAN(+CL)** [65], **DF-GAN** [54], **DM-GAN** [72], **DF-GAN** [54] and, **AttnGAN** [64].

4.5. Human evaluation results

Human evaluation with previous methods is provided in Tab. 4.6. In each instance, human evaluators are required to choose between two images generated by the two models being compared. The two models are compared in three aspects: (i) image quality, (ii) photorealism (which image appears more real), and (iii) text alignment (which image best matches the text). Each question is surveyed using 500 image pairs, where 5 different evaluators answer each question, amounting to 2500 instances per question for a given comparison. We compare our 256×256 model with our re-implementation of DALL-E [45] and CogView's [12] 256×256 model. CogView's 512×512 model is compared with our corresponding model. Results are presented as a percentage of majority votes in favor of our method when comparing between a certain model and ours. Compared with the three methods, ours achieves significantly higher favorability in all aspects.

4.6. FID comparison

FID is calculated over a subset of $30k$ images generated from the MS-COCO validation set text prompts with no re-ranking, and provided in Tab. 4.6. The evaluated models are divided into two groups: trained with and without (denoted as filtered) the MS-COCO training set. In both scenarios our model achieves the lowest FID. In addition, we provide a loose practical lower-bound (denoted as ground-truth), calculated between the training and validation subsets of MS-COCO. As FID results are approaching small numbers, it is interesting to get an idea of a possible practical lower-bound.

4.7. Generating out of distribution

Methods that rely on text inputs only are more confined to generate within the training distribution, as demonstrated by [41]. Unusual objects and scenarios can be challenging to generate, as certain objects are strongly correlated with specific structures, such as cats with four legs, or cars with round wheels. The same is true for scenarios. "A mouse hunting a lion" is most likely not a scenario easily found within the dataset. By conditioning on scenes in the form of simple sketches, we are able to attend to these uncommon objects and scenarios, as demonstrated in Fig. 3, despite the fact that some objects do not exist as categories in our scene (mouse, lion). We solve the category gap by using categories that may be close in certain aspects (elephant instead of mouse, cat instead of lion). In practice, for non-existent categories, several categories could be used instead.

4.8. Scene controllability

Samples are provided in Fig. 1, 3, 4, 5 and in the appendix with both our 256×256 and 512×512 models. In addition to generating high fidelity images from text only, we demonstrate the applicability of scene-wise image control and maintaining consistency between generations.

Model	FID↓	FID↓ (filt.)	Image quality	Photo-realism	Text alignment
AttnGAN [64]	35.49	-	-	-	-
DM-GAN [72]	32.64	-	-	-	-
DF-GAN [54]	21.42	-	-	-	-
DM-GAN+CL [65]	20.79	-	-	-	-
XMC-GAN [67]	9.33	-	-	-	-
DALL-E [45]	-	34.60	81.8%	81.0%	65.9%
CogView$_{256}$ [12]	-	32.20	92.2%	94.2%	92.2%
CogView$_{512}$ [12]	-	36.53	91.1%	88.2%	87.8%
LAFITE [70]	8.12	26.94	-	-	-
GLIDE [41]	-	12.24	-	-	-
Ours$_{256}$	**7.55**	**11.84**			
Ground-truth	2.47	-	-	-	-

Table 1. Comparison with previous work (FID and human preference). FID is calculated over a subset of $30k$ images generated from the MS-COCO validation set text prompts. When possible, we include models trained with and without (filtered) the MS-COCO training set. In both scenarios our model achieves state of the art results, correlating with visual samples and human evaluation. We add a loose practical lower-bound (denoted as ground-truth), calculated between the training and validation subsets of MS-COCO. Human evaluation is shown as a percentage of majority votes in favor of our method when comparing between a certain model and ours.

Model	FID↓	Image quality	Photo-realism	Text alignment
Base	18.01	-	-	-
+Scene tokens	19.16	57.3%	65.3%	58.3%
+Face-aware	14.45	63.6%	59.8%	57.4%
+**CF**	**7.55**	76.8%	66.8%	66.8%
+**Obj-aware$_{512}$**	8.70	62.0%	53.5%	52.2%
+CF with scene input	4.69	-	-	-

Table 2. Ablation study (FID and human preference). FID is calculated over a subset of $30k$ images generated from the MS-COCO validation set text prompts. Human evaluation is shown as a percentage of majority votes in favor of the added element compared to the previous model.

4.9. Scene editing and anchoring

Rather than editing certain regions of images as demonstrated by [45], we introduce new capabilities of generating images from existing or edited scenes. In Fig. 4, two scenarios are considered. In both scenarios the semantic segmentation is extracted from an input image, and used to re-generate an image conditioned on the input text. In the top row, the scene is edited, replacing the 'sky' and 'tree' categories with 'sea', and the 'grass' category with 'sand', resulting in a generated image adhering to the new scene. A simple sketch of a giant dog is added to the scene in the bottom row, resulting in a generated image corresponding to the new scene without any change in text.

Fig. 5 demonstrates the ability to generate new interpretations of existing images and scenes. After extracting the semantic segmentation from a given image, we re-generate the image conditioned on the input scene and edited text.

4.10. Storytelling through controllability

To demonstrate the applicability of harnessing scene control for story illustrations, we wrote a children story, and illustrated it using our method. The main advantages of using simple sketches as additional inputs in this case, are (i) that authors can translate their ideas into paintings or realistic images, while being less susceptible to the "randomness" of text-to-image generation, and (ii) improved consistency between generation. We provide a short video of the story and process.

4.11. Ablation study

An ablation study of human preference and FID is provided in Tab. 4.11 to assess the effectiveness of our different contributions. Settings in both studies are similar to the comparison made with previous work (Sec. 4.3). Each row corresponds to a model trained with the additional element, compared with the model without that specific addition for human preference. We note that while the lowest FID is attained by the 256×256 model, human preference favors the 512×512 model with object-aware training, particularly in quality. Furthermore, we re-examine the FID of the best model, where the scene is given as an additional input, to gain a better notion of the gap from the lower-bound (Tab. 4.6).

5. Conclusion

The text-to-image domain has witnessed a plethora of novel methods aimed at improving the general quality and adherence to text of generated images. While some methods propose image editing techniques, progress is not often directed towards enabling new forms of human creativity and experiences. We attempt to progress text-to-image generation towards a more interactive experience, where people can perceive more control over the generated outputs, thus enable real-world applications such as storytelling. In addition to improving the general image quality, we focus on improving key image aspects we deem significant in human perception, such as faces and salient objects, resulting in higher favorability of our method in human evaluations and objective metrics.

References

[1] Speakers Give Sound Advice. Syracuse post standard. *March*, 28:18, 1911. 1

[2] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. 8

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 7

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 5

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2017. 6

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 8

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 2

[9] Katherine Crowson. *Classifier Free Guidance for Autoregressive Transformers*, 2021b. 7

[10] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 8

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[12] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 4, 8, 9

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2, 3, 4, 5, 6, 7, 13

[14] Marie-Madeleine Fourcade. *L'Arche de Noé: réseau Alliance, 1940-1945*. Plon, 1968. 1

[15] Oran Gafni, Oron Ashual, and Lior Wolf. Single-shot freestyle dance reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 882–891, 2021. 4

[16] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9378–9387, 2019. 4, 13

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[18] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471. PMLR, 2015. 4

[19] 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. 8

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 7, 8

[23] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7986–7994, 2018. 2

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 4

[25] Turgenev Ivan. *Fathers and Sons*. Pandora's Box, 2017. 1

[26] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[27] Horst Woldemar Janson, Anthony F Janson, and Max Marmor. *History of art*. Thames and Hudson London, 1991. 1

[28] Tilke Judd, Frédo Durand, and Antonio Torralba. *A benchmark of computational models of saliency to predict human fixations*, 2012. 2

[29] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 13

[32] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 7

[34] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[35] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[37] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 8

[38] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. 4

[39] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 4

[40] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 4

[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 3, 4, 5, 8, 9

[42] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2, 4

[43] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 2

[44] 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. 2, 4, 7

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2, 3, 4, 8, 9

[46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. *Zero-shot text-to-image generation (ICML spotlight)*, 2021. 2

[47] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3

[48] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 4

[49] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 8

[50] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 7

[51] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 8

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2, 8

[54] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020. 4, 8, 9

[55] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 8

[56] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021. 2

[57] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020. 2

[58] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016. 2

[59] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 3

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[61] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 4

[62] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2, 4

[63] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[64] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 4, 8, 9

[65] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021. 4, 8, 9

[66] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J Zelinsky, and Tamara L Berg. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746, 2013. 2

[67] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–842, 2021. 3, 4, 8, 9

[68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4, 13

[69] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. M6-ufc: Unifying multi-modal controls for conditional image synthesis. *arXiv preprint arXiv:2105.14211*, 2021. 2

[70] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*, 2021. 4, 8, 9

[71] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 4

[72] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 4, 8, 9

A. Additional implementation details

A.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:

$$\alpha_{cat} = \begin{cases} 20, & \text{if cat} \in [154, ..., 158] \\ 1, & \text{otherwise,} \end{cases} \tag{4}$$

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

A.2. 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 $1\times1, 7\times7, 28\times28, 56\times56, 128\times128$, 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 [16]. The per-layer normalizing hyperparameter for the object-aware loss, α_o^l were taken from the work of [13], based on LPIPS [68].

A.3. 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 [31] 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.

B. Additional samples

Additional samples generated from challenging text inputs are provided in Figs. 7-8, while samples generated from text and scene inputs are provided in Figs. 9-12. 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.

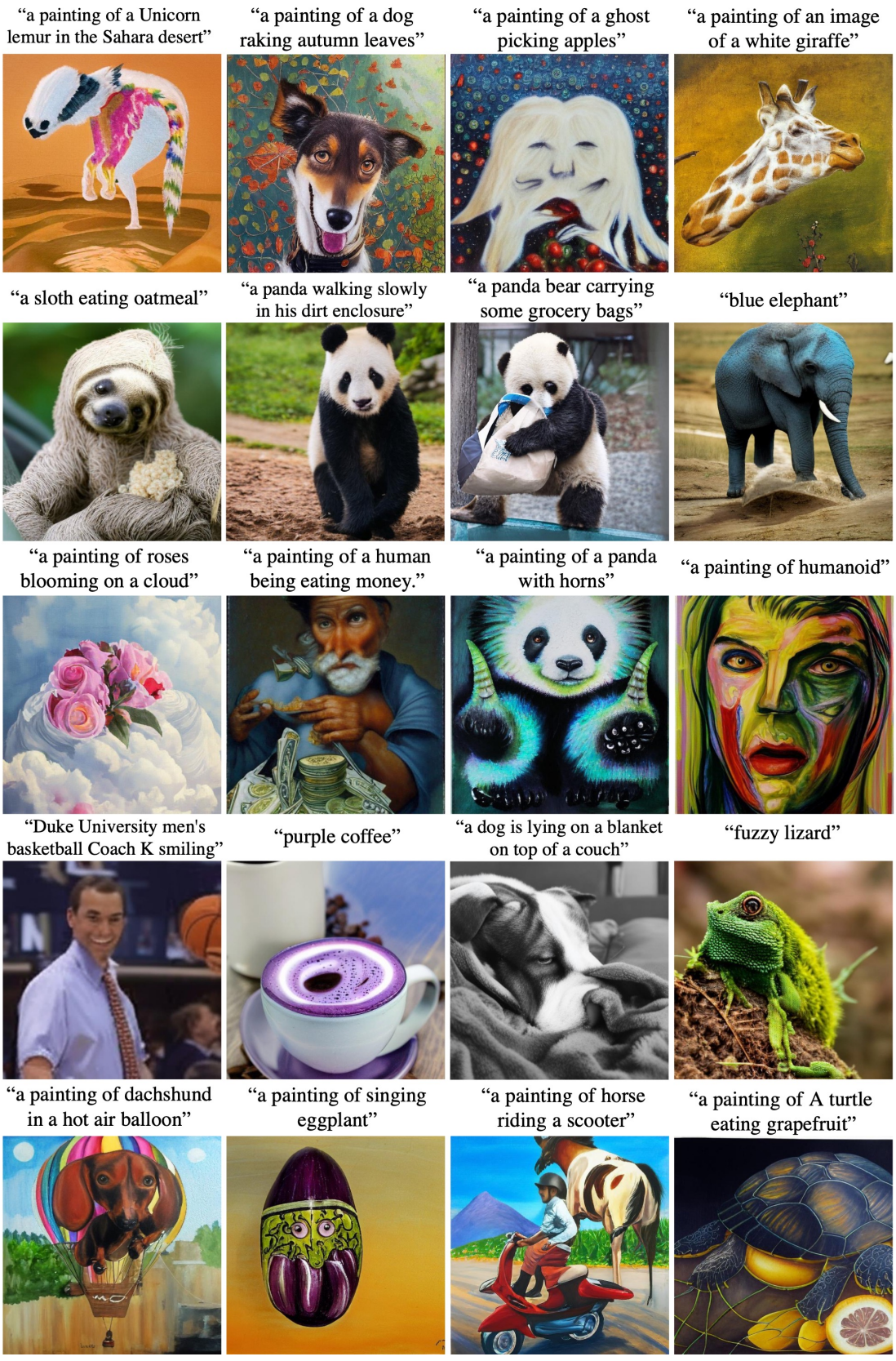
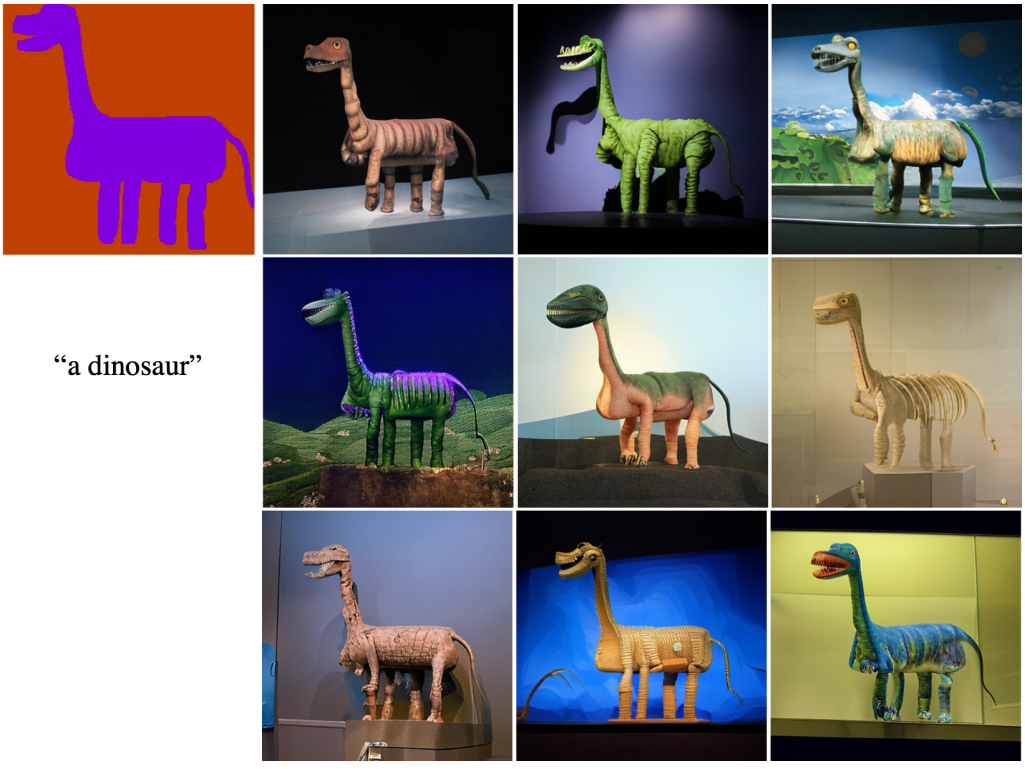


Figure 7. Additional samples generated from challenging text inputs.

Figure 8. Additional samples generated from challenging text inputs.

(a) (b)

Figure 9. Additional samples generated (b) from text and segmentation inputs (a).



(a) (b)

Figure 10. Additional samples generated (b) from text and segmentation inputs (a).

(a) (b)

Figure 11. Additional samples generated (b) from text and segmentation inputs (a).

(a) (b)

Figure 12. Additional samples generated (b) from text and segmentation inputs (a).