A. Derivation of Shifted Diffusion

Recall that we define

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) \coloneqq \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\,\mathbf{z}_{t-1} + \mathbf{s}_t, \beta_t \mathbf{\Sigma}),$$

we use deduction to prove that

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sum_{i=1}^{t} \mathbf{s}_i \sqrt{\bar{\alpha}_t/\bar{\alpha}_i}, (1-\bar{\alpha}_t)\mathbf{\Sigma}). \tag{8}$$

When $t = 1$, we know that

$$q(\mathbf{z}_1 \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_1; \sqrt{1-\beta_1}\,\mathbf{z}_0 + \mathbf{s}_1, \beta_1 \mathbf{\Sigma}).$$

We can re-write

$$\begin{aligned}\sqrt{1-\beta_1}\,\mathbf{z}_0 + \mathbf{s}_1 &= \sqrt{1-\beta_1}\,\mathbf{z}_0 + \sum_{i=1}^{1} \mathbf{s}_i \sqrt{\bar{\alpha}_1/\bar{\alpha}_i} \\ &= \sqrt{\bar{\alpha}_1}\,\mathbf{z}_0 + \sum_{i=1}^{1} \mathbf{s}_i \sqrt{\bar{\alpha}_1/\bar{\alpha}_i}\end{aligned}$$

and

$$\beta_1 \mathbf{\Sigma} = \{1 - (1-\beta_1)\}\mathbf{\Sigma} = (1-\bar{\alpha}_1)\mathbf{\Sigma},$$

which means (8) holds when $t = 1$.

Now we prove the case when $t > 1$. Assume Equation 8 holds for time $t = s$, which means

$$q(\mathbf{z}_s \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_s; \sqrt{\bar{\alpha}_s}\,\mathbf{z}_0 + \sum_{i=1}^{s} \mathbf{s}_i \sqrt{\bar{\alpha}_s/\bar{\alpha}_i}, (1-\bar{\alpha}_s)\mathbf{\Sigma}).$$

We also know that

$$q(\mathbf{z}_{s+1} \mid \mathbf{z}_s) \coloneqq \mathcal{N}(\mathbf{z}_{s+1}; \sqrt{1-\beta_s}\,\mathbf{z}_s + \mathbf{s}_{s+1}, \beta_s \mathbf{\Sigma}),$$

by the property of Gaussian distribution [1], we know that $q(\mathbf{z}_{s+1} \mid \mathbf{z}_0)$ is also a Gaussian distribution with mean

$$\begin{aligned}&(\sqrt{\bar{\alpha}_s}\,\mathbf{z}_0 + \sum_{i=1}^{s} \mathbf{s}_i \sqrt{\bar{\alpha}_s/\bar{\alpha}_i})\sqrt{1-\beta_{s+1}} + \mathbf{s}_{s+1} \\ =&(\sqrt{\bar{\alpha}_s}\,\mathbf{z}_0 + \sum_{i=1}^{s} \mathbf{s}_i \sqrt{\bar{\alpha}_s/\bar{\alpha}_i})\sqrt{1-\beta_{s+1}} + \mathbf{s}_{s+1}\sqrt{\bar{\alpha}_{t+1}/\bar{\alpha}_{t+1}} \\ =&\sqrt{\bar{\alpha}_{s+1}}\,\mathbf{z}_0 + \sum_{i=1}^{s} \mathbf{s}_i \sqrt{\bar{\alpha}_{s+1}/\bar{\alpha}_i} + \mathbf{s}_{s+1}\sqrt{\bar{\alpha}_{t+1}/\bar{\alpha}_{t+1}} \\ =&\sqrt{\bar{\alpha}_{s+1}}\,\mathbf{z}_0 + \sum_{i=1}^{s+1} \mathbf{s}_i \sqrt{\bar{\alpha}_{s+1}/\bar{\alpha}_i},\end{aligned}$$

and covariance matrix (a diagonal matrix)

$$\begin{aligned}&\beta_{s+1}\mathbf{\Sigma} + \sqrt{1-\beta_{s+1}}(1-\bar{\alpha}_s)\mathbf{\Sigma}\sqrt{1-\beta_{s+1}} \\ =&\beta_{s+1}\mathbf{\Sigma} + (1-\beta_{s+1})(1-\bar{\alpha}_s)\mathbf{\Sigma} \\ =&\mathbf{\Sigma}\{\beta_{s+1} + (1-\beta_{s+1})(1-\bar{\alpha}_s)\} \\ =&\mathbf{\Sigma}(\beta_{s+1} + 1 - \beta_{s+1} - \bar{\alpha}_s + \beta_{s+1}\bar{\alpha}_s) \\ =&\mathbf{\Sigma}(1 - \bar{\alpha}_s + \beta_{s+1}\bar{\alpha}_s) \\ =&\mathbf{\Sigma}\{1 - (1-\beta_{s+1})\bar{\alpha}_s\} \\ =&\mathbf{\Sigma}(1-\bar{\alpha}_{s+1}),\end{aligned}$$

which means Equation 8 holds for $t = s + 1$ when it holds for $t = s$. By deduction, we know that since it holds for $t = 1$, it holds for any integer.

Specifically, if we choose $\mathbf{s}_t = (1 - \sqrt{1 - \beta_t})\boldsymbol{\mu}$, we get

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + (1 - \sqrt{\bar{\alpha}_t})\boldsymbol{\mu}, (1 - \bar{\alpha}_t)\boldsymbol{\Sigma}),$$

because

$$\begin{aligned}
&\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sum_{i=1}^{t} \mathbf{s}_i \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_i}} \\
=&\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sum_{i=1}^{t} (1 - \sqrt{1 - \beta_i})\boldsymbol{\mu} \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_i}} \\
=&\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \boldsymbol{\mu} \sum_{i=1}^{t} (1 - \sqrt{1 - \beta_i}) \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_i}} \\
=&\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \boldsymbol{\mu} (\sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_t}} - \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}} + \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}} - \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-2}}} + \ldots + \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_2}} - \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_1}} + \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_1}} - \sqrt{\bar{\alpha}_t}) \\
=&\sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + (1 - \sqrt{\bar{\alpha}_t})\boldsymbol{\mu}.
\end{aligned}$$

We are now able to get the closed-form expression of posterior $q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0)$. We know that

$$\begin{aligned}
q(\mathbf{z}_t \mid \mathbf{z}_0) &= \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + (1 - \sqrt{\bar{\alpha}_t})\boldsymbol{\mu}, (1 - \bar{\alpha}_t)\boldsymbol{\Sigma}) \\
q(\mathbf{z}_{t-1} \mid \mathbf{z}_0) &= \mathcal{N}(\mathbf{z}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\,\mathbf{z}_0 + (1 - \sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\mu}, (1 - \bar{\alpha}_{t-1})\boldsymbol{\Sigma}) \\
q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) &= \mathcal{N}(\mathbf{z}_t; \sqrt{1 - \beta_t}\,\mathbf{z}_{t-1} + (1 - \sqrt{1 - \beta_t})\boldsymbol{\mu}, \beta_t\boldsymbol{\Sigma}).
\end{aligned}$$

From [1], we know that

$$q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\nu}, \boldsymbol{\Lambda})$$

where

$$\begin{aligned}
\boldsymbol{\nu} &= \boldsymbol{\Lambda}\{\sqrt{1 - \beta_t}\{\mathbf{z}_t - (1 - \sqrt{1 - \beta_t})\boldsymbol{\mu}\}\boldsymbol{\Sigma}^{-1}/\beta_t + \{\sqrt{\bar{\alpha}_{t-1}}\,\mathbf{z}_0 + (1 - \sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\mu}\}\boldsymbol{\Sigma}^{-1}/(1 - \bar{\alpha}_{t-1})\}, \\
\boldsymbol{\Lambda} &= \{\boldsymbol{\Sigma}^{-1}/(1 - \bar{\alpha}_{t-1}) + (1 - \beta_t)\boldsymbol{\Sigma}^{-1}/\beta_t\}^{-1}.
\end{aligned}$$

Equation (3) can be obtained by simple derivation.

B. More Experimental Results

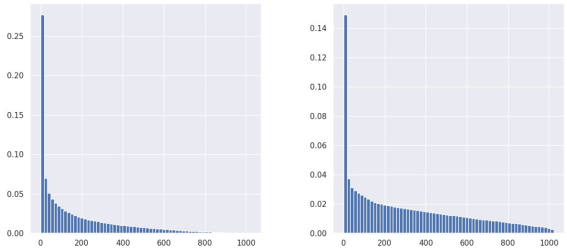
Language-free text-to-image generation We provide more results on language-free text-to-image generation in Table 3. Results in Table 3 are models trained from scratch on corresponding datasets. Results of MM-CelebA-HQ in [34] is based on a model which is first pre-trained on FFHQ dataset [10]. For fair comparison, we use the code provided by the authors and train Lafite-2 from scratch on MM-CelebA-HQ, without pre-training on FFHQ.

	MS-COCO		CUB		LN-COCO		MM-CelebA-HQ	
Methods	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓	IS ↑	FID ↓
Lafite [35]	27.20	18.04	4.32	27.53	18.49	39.85	2.89	32.75
Lafite-2 [34]	31.16	10.26	4.93	16.87	23.18	25.51	2.91	21.89
Corgi (Ours)	**34.14**	10.33	**5.08**	**15.80**	**28.71**	**16.16**	**3.06**	**19.74**

Table 3. Language-free results on MS-COCO, CUB, LN-COCO and MM-CelebA-HQ datasets.

Figure 15. Generated images of mean embeddings from different Gaussian clusters.

Ablation study on shifted diffusion We already know that learn-able Gaussians may lead to higher similarity of generated image embedding to ground-truth embedding, we would like to further compare them. We train 2 shifted diffusion models, with 1024 fixed/learn-able Gaussian distributions. After training, we feed random text captions from the validation set of MS-COCO to the model, and calculate the frequency of each Gaussian being selected by (5). We rank and plot the frequency in Figure 14. Intuitively, when we set $k = 1024$, we are expecting 1024 representative clusters with different semantics, thus we hope every Gaussian has a chance of being selected. However, as we can see in the figure, in the case where 1024 fixed Gaussians are introduced, some Gaussians are never selected which means they contribute nothing to the model. On the contrary, with learned mean and covariance matrix, every Gaussian has a chance of being selected.



(a) 1024 fixed Gaussian (b) 1024 learn-able Gaussian

Figure 14. Frequency of clusters being selected.

A natural question is, are the clusters semantically different? To answer this question, we generate images by directly feeding means of different clusters, which are shown in Figure 15. As we can see, embeddings corresponding to different clusters will lead to generated images that have very different semantics.

C. Implementation Details

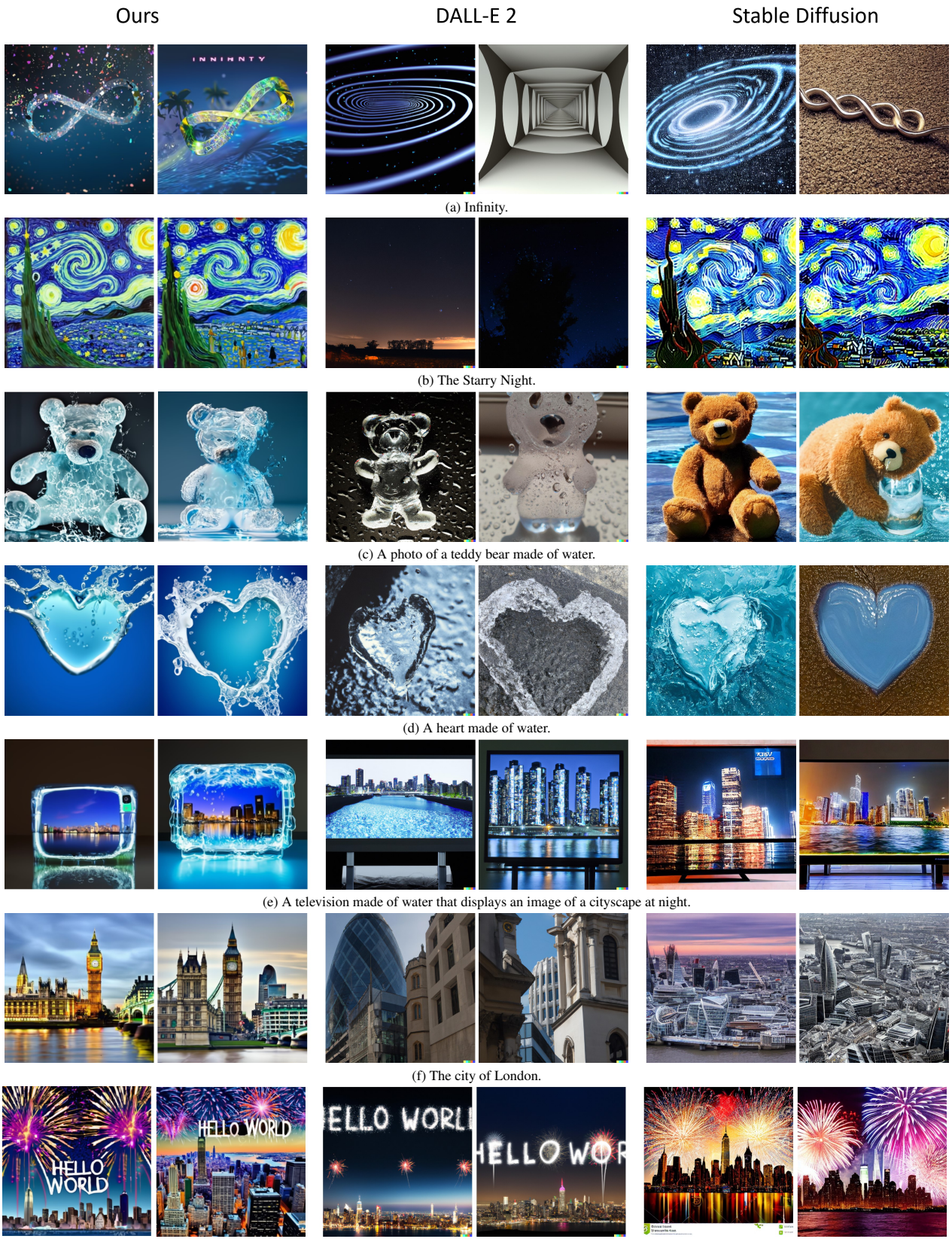
Our prior model trained on 900M dataset is a decoder-only transformer. We set the width, depth, number of attention heads to be 2048, 20, 32 respectively. The model is trained for 500,000 iterations with a batch size of 4096. For the prior model trained on CC15M and baseline prior model in ablation study, we reduce the transformer depth from 20 to 16, while keeping the width and number of attention heads unchanged. The smaller prior models are trained for 40,000 iterations with batch size of 4096 on CC15M dataset. AdamW [14] optimizer with learning rate of 1.2×10^{-4}, $\beta = (0.9, 0.96)$, $\epsilon = 10^{-6}$. We drop the encoded text with probability of 0.1 to enable classifier-free guidance sampling. [9]

Our diffusion-based decoder follows DALL-E 2 [20]. Batch sizes are set to be 2048, 1024, 512 for diffusion models at 64, 256, 1024 resolutions respectively. The models are trained for 1,000,000 iterations. AdamW optimizer with $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$ is used for all three models. Learning rate is set to be 10^{-4} for the model at 1024 resolution, while it is set to be 1.2×10^{-4} for the other two models. We drop the encoded text with probability of 0.1 to enable classifier-free guidance sampling.

Our GAN-based decoder follows the design in [35]. The batch size is set to be 64. Adam optimizer [11] with learning rate 0.0025, $\beta = (0, 0.99)$, $\epsilon = 10^{-8}$ is used for both generator and discriminator.

D. More Generated Examples

We provide more generated examples and comparisons here.



(a) Infinity.

(b) The Starry Night.

(c) A photo of a teddy bear made of water.

(d) A heart made of water.

(e) A television made of water that displays an image of a cityscape at night.

(f) The city of London.

(g) New York Skyline with 'Hello World' written with fireworks on the sky.

Figure 16. Comparison with DALL-E 2 and Stable Diffusion.

(a) A portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. A full moon over the city of Los Angeles is in the background at night.

(b) A photo of an astronaut riding a horse in the forest. There is a river in front of them with water lilies.

(c) A photo of a Ming Dynasty vase on a leather topped table.

(d) A photo of a light bulb in outer space traveling the galaxy with a sailing boat inside the light bulb.

(e) A statue of Abraham Lincoln wearing an opaque and shiny astronaut's helmet. The statue sits on the moon, with the planet Earth in the sky.

(f) Darth Vader playing with raccoon in Mars during sunset.

(g) Ground view of the Great Pyramids and Sphinx on the moon's surface. The back of an astronaut is in the foreground. The planet Earth looms in the sky.

Figure 17. Comparison with DALL-E 2 and Stable Diffusion.

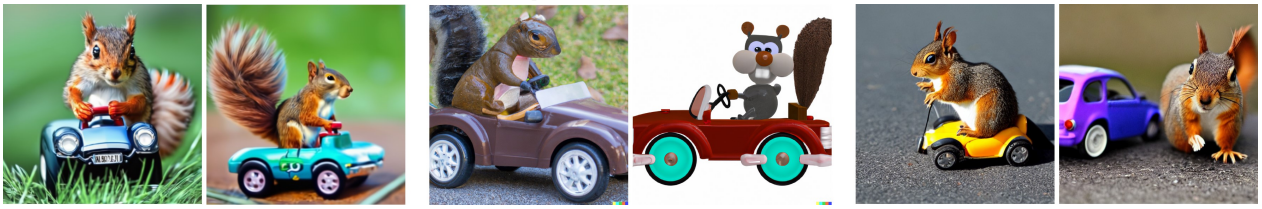
Ours DALL-E 2 Stable Diffusion

(a) A tiger wearing a tuxedo.

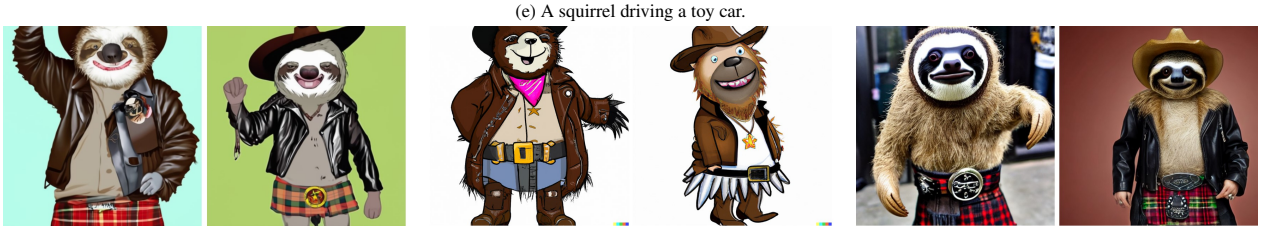
(b) A photograph of an ostrich wearing a fedora and singing soulfully into a microphone.

(c) A photo of llama wearing sunglasses standing on the deck of a spaceship with the Earth in the background.

(d) A teddy bear wearing a motorcycle helmet and cape is riding a motorcycle in Rio de Janeiro with Dois Irmãos in the background.



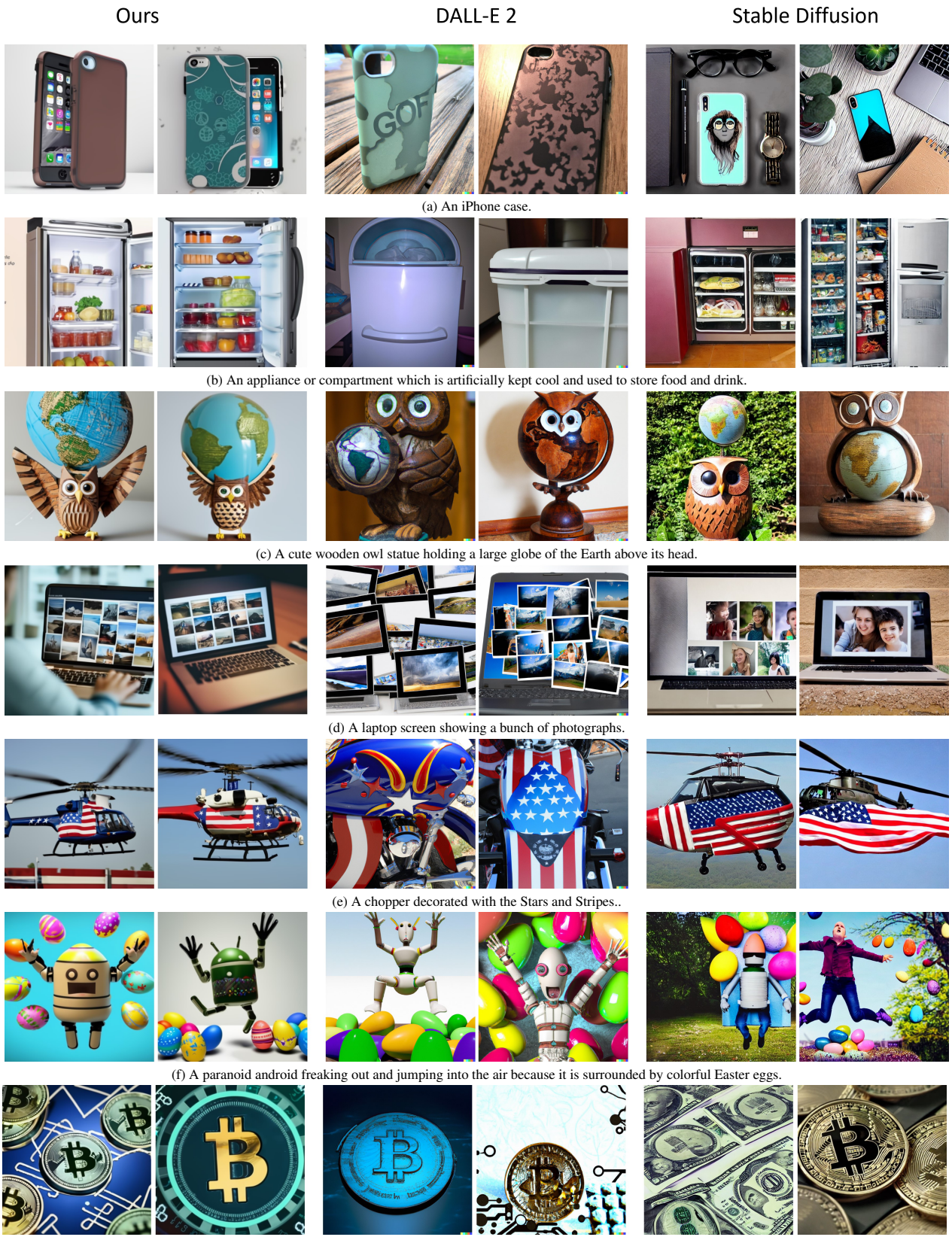
(e) A squirrel driving a toy car.



(f) A smiling sloth wearing a leather jacket, a cowboy hat and a kilt.

(g) A punk rock squirrel in a studded leather jacket shouting into a microphone while standing on a lily pad.

Figure 18. Comparison with DALL-E 2 and Stable Diffusion.



(a) An iPhone case.

(b) An appliance or compartment which is artificially kept cool and used to store food and drink.

(c) A cute wooden owl statue holding a large globe of the Earth above its head.

(d) A laptop screen showing a bunch of photographs.

(e) A chopper decorated with the Stars and Stripes..

(f) A paranoid android freaking out and jumping into the air because it is surrounded by colorful Easter eggs.

(g) A type of digital currency in which a record of transactions is maintained and new units of currency are generated by the computational solution of mathematical problems, and which operates independently of a central bank.

Figure 19. Comparison with DALL-E 2 and Stable Diffusion.

Ours DALL-E 2 Stable Diffusion

(a) A map of the United States made out sushi. It is on a table next to a glass of red wine.

(b) A high resolution photo of a large bowl of ramen. There are several origami boats in the ramen of different colors.

(c) A giant cobra snake made from sushi.

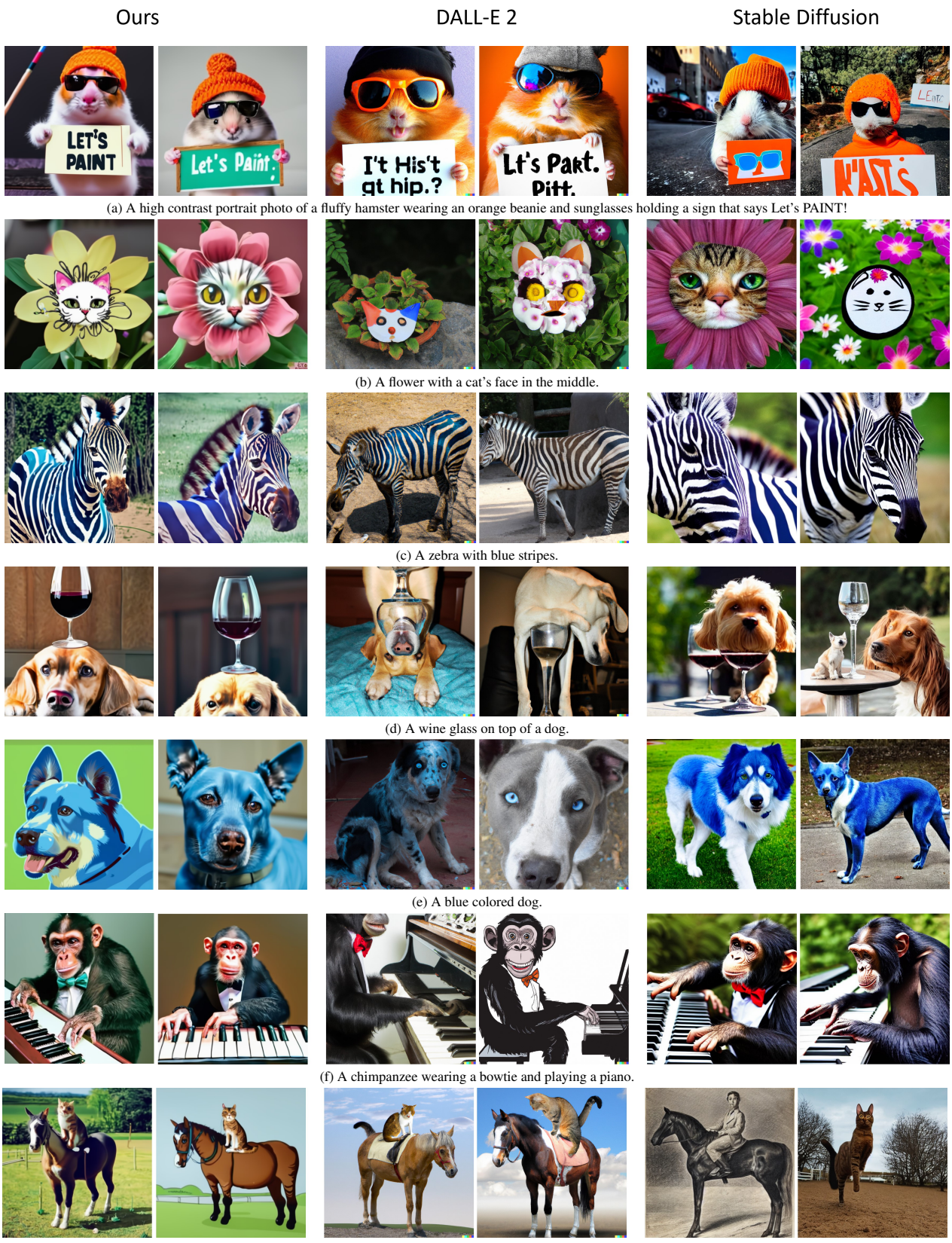
(d) A high resolution photo of a chicken working out in a gym.

(e) Close-up portrait of a smiling businesswoman holding a cell phone, oil painting in the style of Rembrandt.

(f) A young badger delicately sniffing a yellow rose, richly textured oil painting.

(g) A gundam stands tall with its sword raised. A city with tall skyscrapers is in the distance, with a mountain and ocean in the background. A dark moon is in the sky. realistic high-contrast anime illustration.

Figure 20. Comparison with DALL-E 2 and Stable Diffusion.



(a) A high contrast portrait photo of a fluffy hamster wearing an orange beanie and sunglasses holding a sign that says Let's PAINT!

(b) A flower with a cat's face in the middle.

(c) A zebra with blue stripes.

(d) A wine glass on top of a dog.

(e) A blue colored dog.

(f) A chimpanzee wearing a bowtie and playing a piano.

(g) A cat standing on a horse.

Figure 21. Comparison with DALL-E 2 and Stable Diffusion.