Masked Auto-Regressive Variational Acceleration: Fast Inference Makes Practical Reinforcement Learning

Supplementary Material

1. Theory Part

In this section we present a detailed proof of the Gradient Equivalent Theorem. The proof of the Gradient Equivalent Theorem is based on the so-called **Score-projection identity** which was first found in [58] to bridge denoising score matching and denoising auto-encoders. Later the identity is applied for deriving distillation methods based on Fisher divergences [68]. We appreciate the efforts of Zhou et al. [68] and re-write the score-projection identity here without proof. Readers can check Zhou et al. [68] for a complete proof of score-projection identity.

Score-projection identity. Let $u(\cdot, \theta)$ be a vector-valued function, using the notations of the Gradient Equivalent Theorem, under mild conditions, the identity holds:

$$\mathbb{E}_{\substack{\boldsymbol{x}_0 \sim q_{\theta,0} \\ \boldsymbol{x}_t | \boldsymbol{x}_0 \sim p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}} \Big[f(\boldsymbol{x}_t, \boldsymbol{x}_0, \theta) \Big] = 0, \quad \forall \theta \tag{A.1}$$

$$f(\boldsymbol{x}_t, \boldsymbol{x}_0, \theta) = u(\boldsymbol{x}_t, \theta)^T \big\{ s_{q_{\theta,t}}(\boldsymbol{x}_t, c) - \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0, c) \big\}.$$

Next, we turn to prove the Gradient Equivalent Theorem.

Proof of Gradient Equivalent Theorem. We start by applying the chain rule for the total derivative with respect to θ. The function $d(\cdot)$ depends on θ both directly through the score function $s_{q_{\theta,t}}$ and indirectly through the distribution $\boldsymbol{x}_t \sim q_{\theta,t}$ (as $\boldsymbol{x}_t$ depends on $\boldsymbol{x}_0 \sim q_{\theta,0}$). This gives two terms:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_t \sim q_{\theta,t}} \frac{\partial}{\partial \theta} d\left(s_{q_{\theta,t}}(\boldsymbol{x}_t, c) - s_{p_t}(\boldsymbol{x}_t, c) \right) \\
&= \mathbb{E}_{\boldsymbol{x}_t \sim q_{\theta,t}} \Bigg[d'\left(s_{q_{\theta,t}}(\boldsymbol{x}_t, c) - s_{p_t}(\boldsymbol{x}_t, c) \right)^T \frac{\partial}{\partial \theta} s_{q_{\theta,t}}(\boldsymbol{x}_t, c) \\
&+ \frac{\partial}{\partial \boldsymbol{x}_t} d(s_{q_{\theta,t}}(\boldsymbol{x}_t, c) - s_{p_t}(\boldsymbol{x}_t, c)) \frac{\partial \boldsymbol{x}_t}{\partial \theta} \Bigg]
\end{aligned} \tag{A.2}$$

We differentiate (A.1) on both sides w.r.t θ to achieve the equivalent of the first term in (A.2). Since the expectation is zero for all θ, its derivative is also zero. We apply the total derivative (multivariate chain rule), noting that f depends on θ directly, as well as indirectly through $\boldsymbol{x}_0$ and $\boldsymbol{x}_t$ (both depend on θ):

$$0 = \frac{\partial}{\partial \theta} \mathbb{E}_{\substack{\boldsymbol{x}_0 \sim q_{\theta,0} \\ \boldsymbol{x}_t | \boldsymbol{x}_0 \sim p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}} \Big[f(\boldsymbol{x}_t, \boldsymbol{x}_0, \theta) \Big] \tag{A.3}$$

$$= \mathbb{E}\Big[\frac{\partial f}{\partial \boldsymbol{x}_0} \frac{\partial \boldsymbol{x}_0}{\partial \theta} + \frac{\partial f}{\partial \boldsymbol{x}_t} \frac{\partial \boldsymbol{x}_t}{\partial \theta} + \frac{\partial f}{\partial \theta} \Big] \tag{A.4}$$

$$\begin{aligned}
&= \mathbb{E}\Big[\frac{\partial}{\partial \boldsymbol{x}_0} \Big\{ u(x_t, \theta)^T \{ -\nabla_{x_t} \log p_t(x_t|x_0, c) \} \Big\} \frac{\partial x_0}{\partial \theta} \\
&+ \frac{\partial}{\partial \boldsymbol{x}_t} \Big\{ u(\boldsymbol{x}_t, \theta)^T \{ s_{q_{\theta,t}}(x_t, c) - \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|x_0, c) \} \Big\} \frac{\partial \boldsymbol{x}_t}{\partial \theta} \\
&+ \frac{\partial}{\partial \theta} u(\boldsymbol{x}_t, \theta)^T s_{q_{\theta,t}}(\boldsymbol{x}_t, c) + u(\boldsymbol{x}_t, \theta)^T \frac{\partial}{\partial \theta} \{ s_{q_{\theta,t}}(\boldsymbol{x}_t, c) \} \Big]
\end{aligned} \tag{A.5}$$

$$\begin{aligned}
&= \mathbb{E}\Big[\frac{\partial}{\partial \theta} \Big\{ u(x_t, \theta)^T \{ s_{q_{\mathrm{sg}[\theta],t}}(x_t, c) - \nabla_{x_t} \log p_t(x_t|x_0, c) \} \Big\} \\
&+ u(\boldsymbol{x}_t, \theta)^T \frac{\partial}{\partial \theta} \{ s_{q_{\theta,t}}(\boldsymbol{x}_t, c) \Big].
\end{aligned} \tag{A.6}$$

By rearranging the terms in (A.4) and (A.5), and by applying the stop-gradient operator sg to $s_{q_{\theta,t}}$ within the derivative terms, the expression can be consolidated into (A.6). We then rewrite (A.6) as follows:

$$\begin{aligned}
&\mathbb{E}_{x_t \sim q_{\theta,t}} u(\boldsymbol{x}_t, \theta)^T \frac{\partial}{\partial \theta} \{ s_{q_{\theta,t}}(\boldsymbol{x}_t, c) \} \\
&= -\frac{\partial}{\partial \theta} \mathbb{E}_{\substack{\boldsymbol{x}_0 \sim q_{\theta,0} \\ \boldsymbol{x}_t | \boldsymbol{x}_0 \sim p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}} \Big\{ u(x_t, \theta)^T \\
&\{ s_{q_{\mathrm{sg}[\theta],t}}(x_t, c) - \nabla_{x_t} \log p_t(x_t|x_0, c) \} \Big\}
\end{aligned} \tag{A.7}$$

Let $u(\boldsymbol{x}_t, \theta) = d'\left(s_{q_{\mathrm{sg}[\theta],t}}(\boldsymbol{x}_t, c) - s_{p_t}(\boldsymbol{x}_t, c) \right)$ and substitute this specific function u into the identity (A.7):

$$\mathbb{E}_{x_t \sim q_{\theta,t}} \Big[d'\left(s_{q_{\mathrm{sg}[\theta],t}}(\boldsymbol{x}_t, c) - s_{p_t}(\boldsymbol{x}_t, c) \right)^T \frac{\partial}{\partial \theta} \{ s_{q_{\theta,t}}(\boldsymbol{x}_t, c) \} \Big] \tag{A.8}$$

$$= \mathbb{E}_{x_t \sim q_{\theta,t}} \Big[d'\left(s_{q_{\theta,t}}(\boldsymbol{x}_t, c) - s_{p_t}(\boldsymbol{x}_t, c) \right)^T \frac{\partial}{\partial \theta} \{ s_{q_{\theta,t}}(\boldsymbol{x}_t, c) \} \Big] \tag{A.9}$$

$$\begin{aligned}
&= -\frac{\partial}{\partial \theta} \mathbb{E}_{\substack{\boldsymbol{x}_0 \sim q_{\theta,0} \\ \boldsymbol{x}_t | \boldsymbol{x}_0 \sim p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}} \Big\{ d'\left(s_{q_{\mathrm{sg}[\theta],t}}(\boldsymbol{x}_t, c) - s_{p_t}(\boldsymbol{x}_t, c) \right)^T \\
&\{ s_{q_{\mathrm{sg}[\theta],t}}(x_t, c) - \nabla_{x_t} \log p_t(x_t|x_0, c) \} \Big\}
\end{aligned} \tag{A.10}$$

Since θ does not participate in the derivative of $d(\cdot)$, equation (A.8) is equivalent to (A.9). Moreover, θ does not

appear in the differentiation with respect to $\boldsymbol{x}_t$:

$$\mathbb{E}_{x_t\sim q_{\theta,t}}\left[\frac{\partial}{\partial \boldsymbol{x}_t} d(s_{q_{\theta,t}}(\boldsymbol{x}_t,c) - s_{p_t}(\boldsymbol{x}_t,c))\frac{\partial \boldsymbol{x}_t}{\partial \theta}\right] = \\ -\frac{\partial}{\partial \theta}\mathbb{E}_{\substack{\boldsymbol{x}_0\sim q_{\theta,0}\\ \boldsymbol{x}_t|\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}}\left[d\left(s_{q_{\text{sg}[\theta],t}}(\boldsymbol{x}_t,c) - s_{p_t}(\boldsymbol{x}_t,c)\right)\right] \tag{A.11}$$

By combining (A.10) and (A.11) into (A.2), we get exactly the **Gradient Equivalent Theorem** stated in the text. This completes the proof.

2. Implementation Details

In this section, we provide detailed hyperparameter settings and the algorithmic procedure for MARVAL. Our framework consists of two distinct stages: the Guided Score Implicit Matching (GSIM) distillation stage and the Reinforcement Learning (RL) refinement stage.

2.1. Hyperparameter Settings

Model Architectures. The architecture of our one-step generator follows the same configuration of the original MAR models [68].

Hyperparameter	Stage 1: GSIM Distillation
GPUs	8 × NVIDIA A100
Training Epochs	30
Approx. Training Time	$\sim$ 3 days
Teacher Steps (N_{diff})	1000
Student Steps (N_{diff})	1
CFG Scale (w)	1.2
Loss Function	Pseudo-Huber ($r = 10^{-5}$)
Student model Lr	5e-6
Auxiliary model Lr	5e-6
EMA momentum	0.9999
batch size per GPU	64

Table 1. Training Settings for MARVAL Distillation stage.

Training Configurations. The training is conducted on NVIDIA A100 GPUs. We utilize the AdamW optimizer for both stages. The specific hyperparameters for the Distillation and RL stages are provided in Table 1 and Table 2.

During the distillation stage, we set the teacher's full diffusion steps to $N_{diff} = 1000$ and the one-step generator predicts the noise from the fix step 400. Based on our ablation study, we set the Classifier-Free Guidance (CFG) scale $w = 1.2$ for the teacher score, which provides the optimal trade-off between fidelity (FID) and fidelity. For the loss function, we use the Pseudo-Huber distance with $r = 1e{-}5$ to ensure numerical stability.

Hyperparameter	Stage 2: RL Refinement
GPUs	8 × NVIDIA A100
Training Epochs	5
Approx. Training Time	$\sim$ 2 days
AR Loops (N_{AR})	64
Student Steps (N_{diff})	1
Reward Model	Pickscore
Student model Lr	5e-6
EMA momentum	0.9999
batch size per GPU	2

Table 2. Training Settings for RL Refinement stage.

During the RL refinement stage, we use PickScore [22] as the reward model. We generate samples with AR loops(N_{AR})=64 to calculate rewards, ensuring the optimization aligns with the final inference quality and requires less memory costs. Our best results in FID and IS are tested when inferring with $N_{AR} = 128$.

3. Additional Results

This section provides additional qualitative results for the MARVAL-RL-L and MARVAL-RL-H models, as well as extended text-to-image generation samples from our 1-step+RL DC-AR model. The MARVAL results were not included in the main paper for the following reasons:

- First, due to space limitations, we primarily presented results using the MARVAL-RL-B model.
- Second, the qualitative visualizations in the main experiments adopt MARVAL-RL-B because it already delivers strong visual quality while maintaining the fastest inference speed among all model sizes.
- Third, the MARVAL-L and MARVAL-H models achieve relatively strong performance even before RL fine-tuning, making the improvements brought by RL most pronounced and interpretable on MARVAL-B.

For completeness, we include the MARVAL-RL-L and MARVAL-RL-H visualizations in Fig. 1 and Fig. 2, respectively. Compared with MARVAL-RL-B and MARVAL-RL-L, the MARVAL-RL-H samples exhibit significantly richer fine-grained textures, particularly on man-made objects such as buildings, tools, and clothes. This trend provides additional evidence that our method scales effectively: as model capacity increases, the RL fine-tuning yields increasingly detailed and visually faithful results.

Furthermore, to supplement the text-to-image experiments presented in the main text, we provide a broader set of qualitative samples generated by our 1-step+RL DC-AR model in Fig. 3.

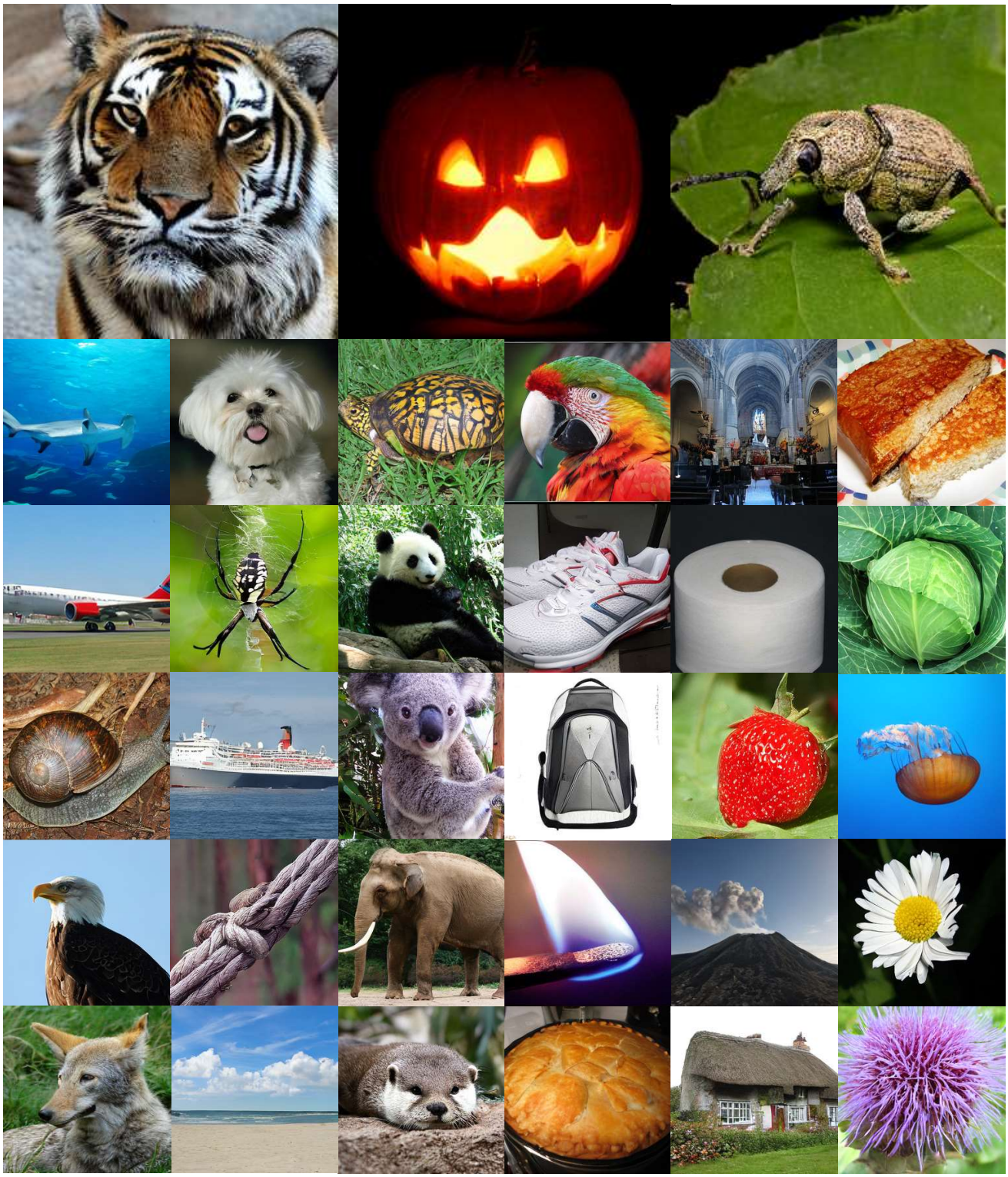


Figure 1. Qualitative results of MARVAL-RL-L.

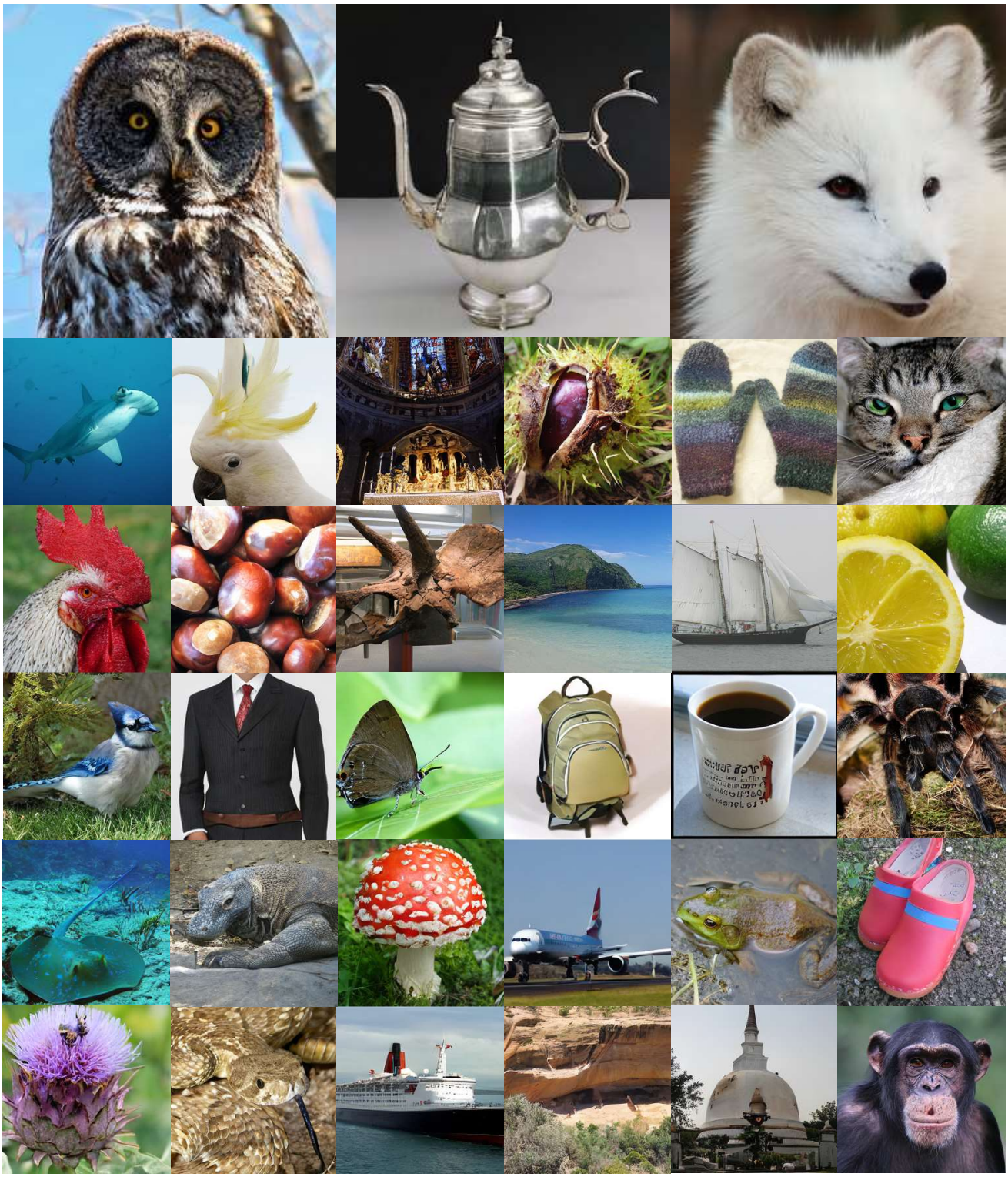


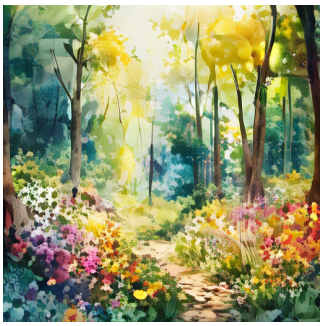
Figure 2. Qualitative results of MARVAL-RL-H.

Cheetah, pastel, cinematic lighting, full body view, stylize, detailed, v4.

The face of god, light, holy.

Fluffy koala bear happy.

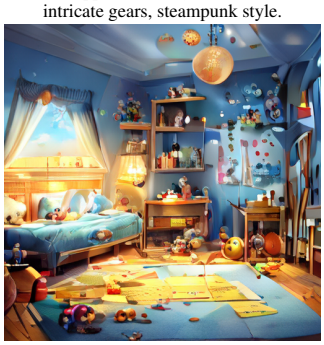
Mechanical hummingbird made of brass and copper, hovering near flowers, intricate gears, steampunk style.



A watercolor of a beautiful forest with flowers, sunny, colorful.

A haunted house on a hill under a full moon.

Professional portrait of an anthropomorphic cat wearing gentleman hat and jacket walking in autumn forest.



Boy's room, children's room, comfortable room with toys, Freelancer game, drawing for a children's storybook, blue room, cartoon.

Starfield, deep sky, realistic, 8k

Dragon made of constellation stars flying across night sky, over mountain landscape.

Skeleton, in the snow, friendly, made with pen and ink and prismacolor shading, blue and white color palet.

An elegant gold and white tikka head jewelry.

A framed renaissance painting of a unicorn on another planet in outer space. oil painting on canvass. rembrandt. 1647.

Christina Hendricks as a Disney princess, Full Torso, standing on a city street, octane render, volumetric lighting.

A painting of a rooster on a white background, a watercolor painting by Sun Kehong, shutterstock contest winner, cloisonnism, detailed painting, behance hd, photoillustration, 8K, UHD.

Starfield, space, 8k, photo, realism, sharp photography, maximum detail, sharp focus, Intricate details, epic, wide shot, highly realistic, cinematic lighting, volumetric lighting, octane render.

Figure 3. More qualitative results of distilled+RL DC-AR text-to-image generation. Our method achieves high generation quality and strong text alignment, demonstrating robust adaptability to prompts of varying lengths.