Consistent3D: Towards Consistent High-Fidelity Text-to-3D Generation with Deterministic Sampling Prior

Supplementary Material

A. Discussion

Diffusion models start by diffusing $p_{\text{data}}(\mathbf{x})$ with a stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x} = \mu(t)\mathbf{x}\mathrm{d}t + \sigma(t)\mathrm{d}\mathbf{w}, \tag{14}$$

where $t \in [0, T]$, $\mu(\cdot)$ and $\sigma(\cdot)$ are the drift and diffusion coefficients respectively, and $\mathbf{w}$ denotes the standard Brownian motion. We denote the distribution of $\mathbf{x}_t$ by $p_t(\mathbf{x})$. A notable characteristic of this SDE is that there exists an Ordinary Differential Equation (ODE), named the Probability Flow (PF) ODE [41], whose solution trajectories, when sampled at time t, adhere to the distribution $p_t(\mathbf{x})$:

$$\mathrm{d}\mathbf{x} = \left[\mu(t)\mathbf{x} - \frac{1}{2}\sigma(t)^2\nabla \log p_t(\mathbf{x})\right]\mathrm{d}t. \tag{15}$$

Due to the above connection between the PF ODE and forward SDE, one can sample along the distribution of the ODE trajectories by first sampling $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$, then adding Gaussian noise to $\mathbf{x}$. This implies that we can effectively sample two solutions on the PF ODE trajectory by first rendering an image $\mathbf{x}_\pi$, followed by perturbing it with Gaussian noise $\boldsymbol{\epsilon}^*$. In this paper, we follow Karras et al. [16] and formulate the forward and reverse process as illustrated in Sec. 3, particularly $\mu(t) = 0$ and $\sigma(t) = \sqrt{2t}$. Thus, the perturbed sample is as follows:

$$\mathbf{x}_{t_1} = \mathbf{x}_\pi + \sigma_{t_1}\boldsymbol{\epsilon}^*, \tag{16}$$

and then computing $\hat{\mathbf{x}}_{t_2}$ using one discretization step of the numerical ODE solver by:

$$\hat{\mathbf{x}}_{t_2} = \mathbf{x}_{t_1} + \frac{\sigma_{t_2} - \sigma_{t_1}}{\sigma_{t_1}}(\mathbf{x}_{t_1} - D_\phi(\mathbf{x}_{t_1}, t_1)). \tag{17}$$

By optimizing the underlying 3D representation $\boldsymbol{\theta}$ to find an optimal $\mathbf{x}_\pi$, we can minimize the distance of the predicted sample from $D_\phi(\cdot)$ given $\mathbf{x}_{t_1}$ and $\hat{\mathbf{x}}_{t_2}$, *i.e.*, minimizing the discrepancy of $D_\phi(\mathbf{x}_{t_1}, t_1)$ and $D_\phi(\hat{\mathbf{x}}_{t_2}, t_2)$ (Eq. (10)). Therefore, we can eventually align $\mathbf{x}_\pi$, $\hat{\mathbf{x}}_{t_2}$ and $\mathbf{x}_{t_1}$ into the same deterministic flow, thus making $\mathbf{x}_\pi$ a realistic data point as it becomes a solution of the ODE sampling flow. Note that although this process introduces a truncation error from numerical ODE solvers, we will prove that our CDS achieves the same accuracy as multi-step sampling approaches in a single-step generative framework in Appendix C, and this error bound is almost optimal since one always needs to discretize the ODE flow to simulate it by diffusion models.

B. Experiments

B.1. Additional Implementation Details

Rendering settings. For each iteration, we randomly select 12 camera poses with an 80% probability of rendering normal maps and a 20% probability of rendering colored images. The field of view (fovy) range is randomly sampled between 30 and 45 degrees, while the azimuth angles will be discussed in the following paragraph. The initial spherical radius is 2.0, and the camera distance is randomly sampled between 1.5 and 2.0. We did not use soft shading, as we found that it significantly slows down the training process. The final rendering resolutions are set to 64×64 and 512×512 for the coarse and fine stages, respectively.

Modified batch uniform azimuth sampling. We observe that vanilla batch uniform azimuth sampling is not suitable for view-dependent prompting selected within {`front`, `side`, `back`} when the batch size is large (*e.g.*, 12). This can lead to the Janus face issue, as the same guidance is shared between different azimuths. We empirically found that splitting the azimuth range into 4 parts according to the prompt and uniform sampling of azimuths within each range can alleviate this problem.

Annealed Classifier-free Guidance. We empirically find that reasonably large CFG weight leads to better details, which accords with the empirical results in DreamFusion [31]. The optimization-based generation framework has a single-step sampling approach that is distinct from the multi-step sampling used in image generation. This single-step sampling framework requires a typically larger CFG to emphasize the details that appear in a single generation, whereas multi-step sampling is able to accumulate details many times over, thus allowing the use of smaller CFG. This framework also differs from other frameworks that require training of LoRA [47], as LoRA plays a similar role to the high CFG values, *i.e.*, providing better orientation in the single-step generation. However, oversaturation problems can occur in the generated images if the CFG values are too high for small time steps. Therefore, we suggest that the CFG value should also vary with the time step schedule. In other words, it should become progressively smaller. In practice, we linearly decrease the CFG value from 50 to 20 with increasing iteration.

Evaluation Settings. For quantitative evaluation, we measure the CLIP R-precision [32] following the practice of DreamFusion [31]. We compare with DreamFusion [31], Magic3D [20] and ProlificDreamer [47] using 40 randomly

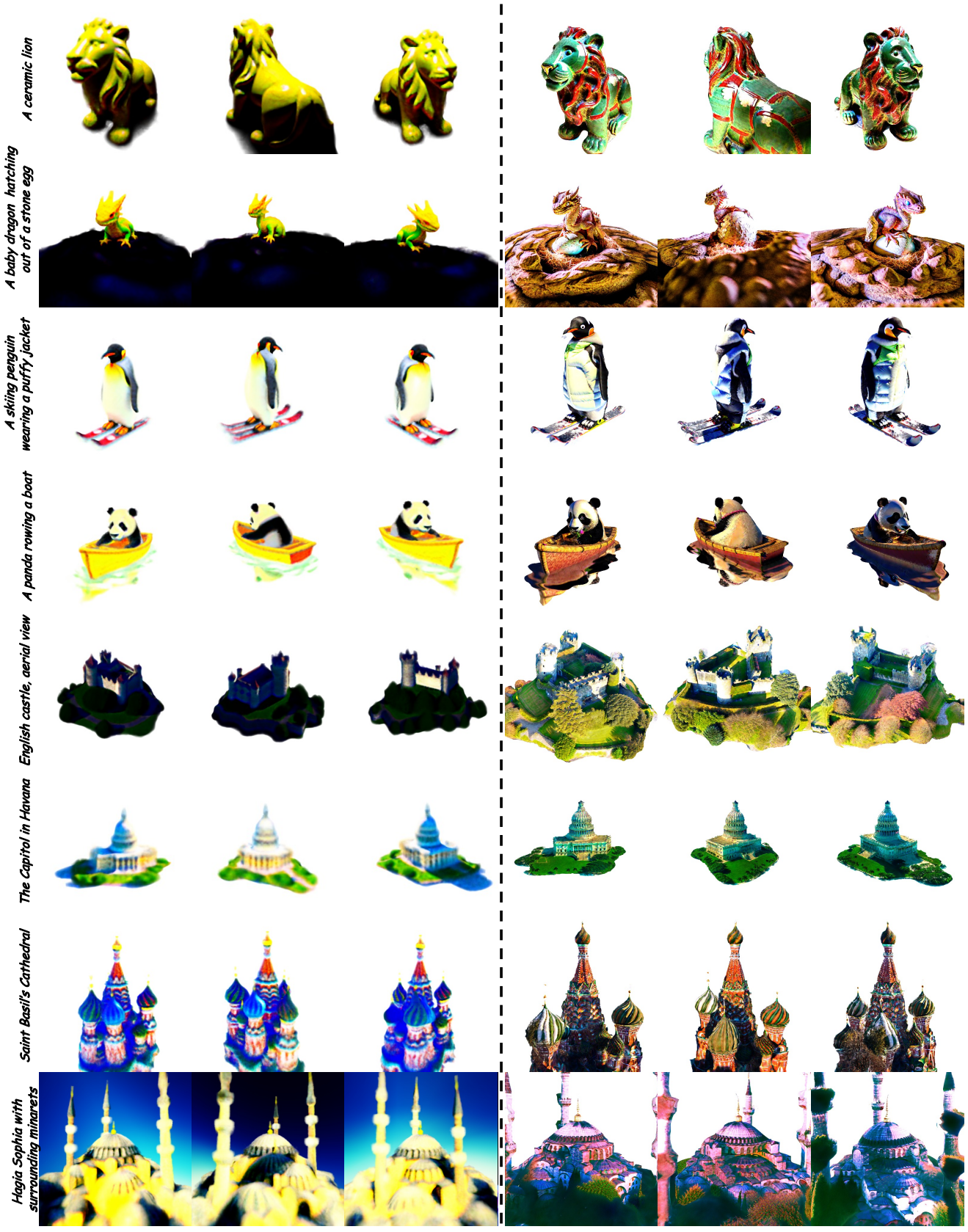

Figure 7. **Qualitative Results** of Two-Stage text-to-3D Generation: NeRF Optimization and Mesh Refinement.

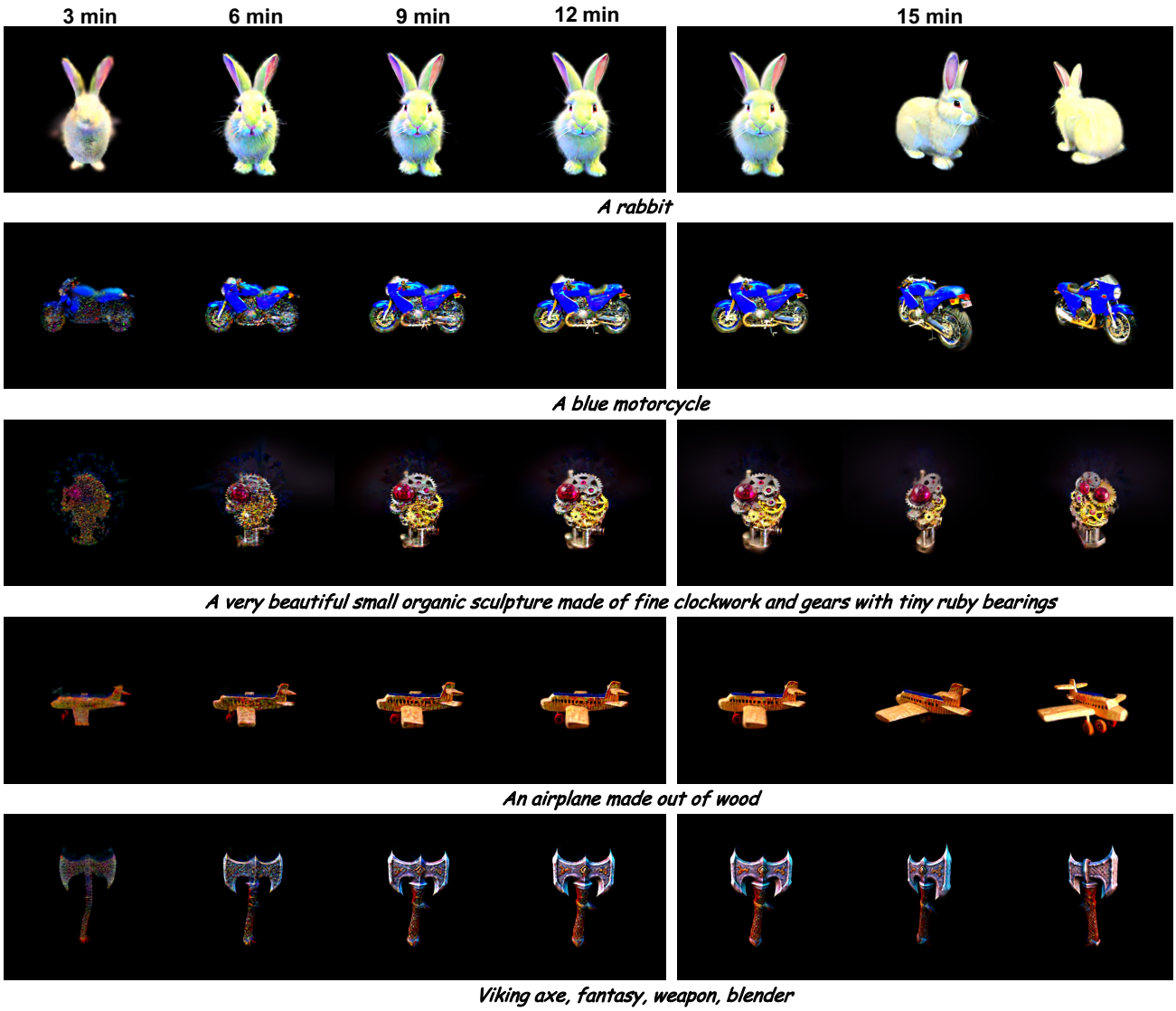

Figure 8. **Text-to-3D Generation using Consistent3D with 3D Gaussian Splatting.** Qualitative results demonstrate that our Consistent3D is capable of producing detailed, high-fidelity 3D models from text prompts in 15 minutes.

selected text prompts from the DreamFusion gallery [1]. The prompts used in our experiments are listed in Tab. 4. We then render 120 views with uniformly sampled azimuth angles and calculate the CLIP R-precision based on each rendered image, and we average the different views for the final metric.

B.2. Additional Results

More qualitative results for both the coarse and fine stages are shown in Fig. 7. We notice that our Consistent3D can generate robust geometry in the coarse stage and then seemly enhance high-frequency and sophisticated details in the fine stage. We conduct the ablation study quantitatively

[1] https://dreamfusion3d.github.io/gallery.html

(a)	(b)	(c)	(d) Ours
0.319	0.325	0.340	**0.348**

Table 2. CLIP R-precision of the ablation study. (a) random time step schedule; (b) predetermined time step schedule; (c) random noise in each iteration; (d) our proposed configuration.

	DF	M3D	PD	Ours
User Prefer. Rate (↑)	15.0%	5.4%	15.0%	**64.6%**
GPT4-V Avg. Rank (↓)	3.0	3.8	1.8	**1.4**

Table 3. **User/AI study.** DF: DreamFusion; M3D: Magic3D; PD: ProlificDreamer.

in Tab. 2, the results also support the superiority of our Consistent3D over others.

We also conduct a user study in Tab. 3 where 50 users select best from multiple choices on 30 generated 3D assets. Moreover, GPT4-V is used to evaluate and rank the generated 3D assets from aesthetic appeal, multi-view consistency and alignment with text prompt. Tab. 3 demonstrates the superiority of the results generated by our Consistent3D from both human and large multi-modal AI perspectives.

B.3. Fast Generation with 3D Gaussian Splatting

Our Consistent3D with Consistency Distillation Sampling is a general text-to-3D generation framework that can be used to create a variety of 3D representations, including 3D Gaussian Splatting [17]. As demonstrated in Fig. 8, our Consistent3D is capable of producing high-fidelity 3D models with intricate details in 15 minutes, which vividly showcases the potential of CDS among different 3D representations (NeRF, Mesh, 3D Gaussian Splatting).

C. Theoretical Proof

Theorem 1. *Assume that the diffusion model $D_\phi(\cdot)$ satisfies the Lipschitz condition. Define $\Delta := \sup|t_1 - t_2|$. For any given camera pose π, if convergence is achieved according to Eq. (10), then there exists a corresponding real image $\mathbf{x}^* \sim p_{data}(\mathbf{x})$ such that*

$$\|\mathbf{x}_\pi - \mathbf{x}^*\|_2 = \mathcal{O}(\Delta), \tag{18}$$

where $\mathbf{x}_\pi = g(\boldsymbol{\theta}, \pi)$ denotes the rendered image for pose π.

Proof. From $\mathcal{L}_{\text{CDS}}(\boldsymbol{\theta}; \pi) = 0$, for any given π and $T \geq t_n > t_{n+1} \geq 0$, it satisfied that

$$D_\phi(\mathbf{x}_{t_n}, t_n, y) \equiv D_\phi(\hat{\mathbf{x}}_{t_{n+1}}, t_{n+1}, y). \tag{19}$$

Let $\boldsymbol{e}_n$ represent the error at t_n, which is defined as:

$$\boldsymbol{e}_n := D_\phi(\mathbf{x}_{t_n}, t_n) - \mathbf{x}^*. \tag{20}$$

We can derive the error at t_{n+1} given the error at t_n:

$$\begin{aligned}
\boldsymbol{e}_n &= D_\phi(\mathbf{x}_{t_n}, t_n) - \mathbf{x}^* \\
&= D_\phi(\hat{\mathbf{x}}_{t_{n+1}}, t_{n+1}) - \mathbf{x}^* \\
&= D_\phi(\hat{\mathbf{x}}_{t_{n+1}}, t_{n+1}) - D_\phi(\mathbf{x}_{t_{n+1}}, t_{n+1}) \\
&\quad + D_\phi(\mathbf{x}_{t_{n+1}}, t_{n+1}) - \mathbf{x}^* \\
&= D_\phi(\hat{\mathbf{x}}_{t_{n+1}}, t_{n+1}) - D_\phi(\mathbf{x}_{t_{n+1}}, t_{n+1}) + \boldsymbol{e}_{n+1}.
\end{aligned}$$

According to the Lipschitz condition, we can further derive

$$\begin{aligned}
\|\boldsymbol{e}_n\| &\leq \|D_\phi(\hat{\mathbf{x}}_{t_{n+1}}, t_{n+1}) - D_\phi(\mathbf{x}_{t_{n+1}}, t_{n+1})\| + \|\boldsymbol{e}_{n+1}\| \\
&\leq L\|\hat{\mathbf{x}}_{t_{n+1}} - \mathbf{x}_{t_{n+1}}\| + \|\boldsymbol{e}_{n+1}\| \\
&\overset{(i)}{=} \|\boldsymbol{e}_{n+1}\| + \mathcal{O}((t_n - t_{n+1})^2),
\end{aligned}$$

where (i) hold according to the local error of Euler solver. Therefore, we can derive the error recursively:

$$\begin{aligned}
\|\boldsymbol{e}_0\| &\leq \sum_{k=0}^{N-1} \mathcal{O}((t_k - t_{k+1})^2) \\
&\leq \sum_{k=0}^{N-1} (t_k - t_{k+1})\mathcal{O}(\Delta) \\
&= \mathcal{O}(\Delta).
\end{aligned}$$

□

Remark. Theorem 1 not only affirms the realistic rendering capabilities of CDS for any camera pose π, but also highlights that CDS achieves an error bound of $\mathcal{O}(\Delta)$, which is the same as multi-step sampling approaches [39, 43]. This shows that CDS is capable of achieving the same accuracy as multi-step approaches in a single-step generative framework, thus demonstrating its efficiency and broad applicability for optimization-based generation.

D. Limitations

Our approach relies on pre-trained diffusion models without 3D priors, and it may sometimes produce less than satisfactory results, especially in complex 3D modeling scenarios. Additionally, pre-trained models may unintentionally transfer unwanted bias from their original training data and parameters into the generated 3D models.

These challenges highlight two critical areas for future research and development in 3D generation. First, there is a pressing need to develop generative models that incorporate robust 3D-centric training. Such models would be better equipped to handle the complexities and nuances inherent in 3D structures. Second, it is essential to devise strategies that effectively identify and neutralize biases transferred from pre-trained models. Addressing these issues will not only improve the accuracy and reliability of 3D generation, but will also ensure the ethical integrity and fairness of the generative process.

ID	Prompt
1	A DSLR photo of a squirrel dressed like a clown
2	A DSLR photo of a tiger made out of yarn
3	A DSLR photo of the Imperial State Crown of England
4	A beagle in a detective's outfit
5	A blue motorcycle
6	A blue poison-dart frog sitting on a water lily
7	A cat with a mullet
8	A ceramic lion
9	A highly detailed sand castle
10	A lemur taking notes in a journal
11	A panda rowing a boat
12	A panda wearing a necktie and sitting in an office chair
13	A photo of a skiing penguin wearing a puffy jacket, highly realistic DSLR photo
14	A photo of a tiger dressed as a doctor, highly realistic DSLR photo
15	A photo of the Ironman, highly realistic DSLR photo
16	A rabbit, animated movie character, high detail 3D model
17	A silver platter piled high with fruits
18	A squirrel dressed like Henry VIII king of England
19	A tarantula, highly detailed
20	A tiger wearing a tuxedo
21	A wide-angle DSLR photo of a squirrel in samurai armor wielding a katana
22	A zoomed-out DSLR photo of a 3D model of an adorable cottage with a thatched roof
23	A zoomed-out DSLR photo of a human skeleton relaxing in a lounge chair
24	A zoomed-out DSLR photo of a model of a house in Tudor style
25	A zoomed-out DSLR photo of a panda throwing wads of cash into the air
26	An astronaut is riding a horse
27	The Great Wall, aerial view
28	Michelangelo style statue of dog reading news on a cellphone
29	A DSLR photo of a baby dragon hatching out of a stone egg
30	A DSLR photo of a bald eagle
31	A DSLR photo of a bear dressed in medieval armor
32	A DSLR photo of a humanoid robot holding a human brain
33	A DSLR photo of a peacock on a surfboard
34	A DSLR photo of a red pickup truck driving across a stream
35	A DSLR photo of a Shiba Inu playing golf wearing tartan golf clothes and hat
36	A 20-sided die made out of glass.
37	A DSLR photo of Mount Fuji, aerial view.
38	An old vintage car.
39	A delicious hamburger.
40	A cute steampunk elephant.

Table 4. **Prompt library** for quantitative results.