Appendix of "Connecting Consistency Distillation to Score Distillation for Text-to-3D Generation"

Zongrui Li[1,2] †, Minghui Hu[2] †, Qian Zheng[3,4] ✉, and Xudong Jiang[2]

[1] Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University
[2] School of Electrical and Electronic Engineering, Nanyang Technological University
[3] College of Computer Science and Technology, Zhejiang University
[4] The State Key Lab of Brain-Machine Intelligence, Zhejiang University
{ZONGRUI001, e200008, EXDJiang}@ntu.edu.sg, qianzheng@zju.edu.cn

In this appendix,

1. we provide some discussions about the limitations of our method and the failure cases;
2. we show an example question for the user study in Sec. 8;
3. we visualize the accumulated brightness during optimization in Sec. 9;
4. we provide a detailed proof of Lemma 1 in Sec. 10 that clarifies the effectiveness of the proposed Compact Consistency (CC) loss;
5. we show more ablation studies regarding CC and Conditional Guidance (CG) loss in Sec. 11; we also show the effect of high order ODE solver in this section;
6. we showcase more results generated by the proposed Guided Consistency Sampling (GCS) in Sec. 12.

7 Limitations

The proposed GCS exhibits shortcomings in addressing Janus issues, which we attribute to the inherent inconsistencies in pre-trained models under multi-view conditions. Also, it doesn't perform well on compositional prompts. Moreover, our approach necessitates extensive training in the GS model, ranging from 3000 to 5000 epochs, translating to 35-60 minutes for rendering a single object. These limitations show a significant gap in practical application.

Failure case I: Janus Problem. Despite proposing multiple strategies to improve the overall generation quality, our method still suffers from the Janus problem and occasionally generates multi-view inconsistent results (Fig. 7). A solution is to change the random seed or give a better initialization.

Fig. 7: Failure case I. The 3D asset is affected by the Janus problem.

† Equal contribution. ✉ Corresponding author.

Failure case II: Compositional Prompt. Our method performs poorly in generating high-quality compositional objects (Fig. 8), possibly due to a bad initialization with insufficient Gaussian points to model different components or a lack of relevant techniques to combine those components correctly. We regard improving this as a future work.

Fig. 8: Failure case II. The 3D assets are generated based on a compositional prompt.

8 Example Question

We show an example question from our user study. We use Google Forms to collect the volunteers' responses.

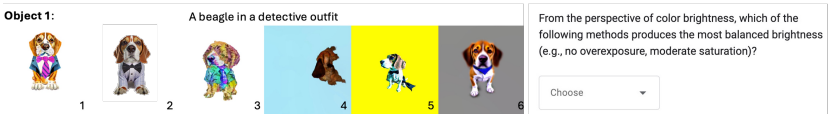

Fig. 9: Example questions used in the user study.

9 Visualization of Accumulated Brightness

We show an example of accumulated brightness during the generation process of the proposed method without BEG in Fig. 10.

Fig. 10: An illustration of brightness accumulation. The highlight point (red box) in early-stage training will be kept and spread to its surroundings, causing over-saturation.

10 Proof of Lemma 1

Lemma 1 ([1, 4, 6, 8]). *Let $\Delta t = \max\{|\delta_k|\}$, $k \in [0, ..., n_s]$, where n_s is the index of δ at time step s, and $F_\theta(\cdot,\cdot)$ is the origin prediction function grounded on the empirical PF-ODE. Assume F_θ satisfies the Lipschitz condition, if there is a $\mathbf{x}_\pi$ satisfying $\mathcal{L}_{\mathrm{CC}}(\xi) = 0$, given an image $\mathbf{x}_0 \sim p_{data}(\mathbf{x})$, for any $t, s, e \in [\![0, ..., T]\!]$ with $t > s > e$, we have:*

$$\sup_{t,e,\mathbf{x}_\pi} \|\hat{\mathbf{x}}_{\{\pi,e\}}, \hat{\mathbf{x}}_{\{0,e\}}\|_2 = \mathcal{O}\left((\Delta t)^p\right)(T-e), \tag{26}$$

$\hat{\mathbf{x}}_{\{0,e\}}$ is the distribution of $\mathbf{x}$ diffused to time e, p is the order of the ODE solver.

Proof. *Given $\mathcal{L}_{\mathrm{CC}}(\xi) = 0$, for any π, t, s, and e, we have:*

$$G_\theta\left(\hat{\mathbf{x}}_{\{\pi,t\}}; t, e, y\right) \equiv G_\theta\left(\hat{\mathbf{x}}_{\{\pi,s\}}; s, e, y\right). \tag{27}$$

Defining $\hat{\mathbf{x}}_{\{\pi,t_k\}} = \mathbf{x}_{t_k}$ for simplicity, Eq. (27) can be rewritten as:

$$G_\theta\left(\mathbf{x}_s; s, e, y\right) \equiv G_\theta\left(\mathbf{x}_t; t, e, y\right), \tag{28}$$

where, $\mathbf{x}_t$ is obtained by $\hat{\mathbf{x}}_s$ through DDIM inversion. For a more general expression, we represent distillation error [6] at $t_k > e$ as:

$$e_e^{t_k} = G_\theta(\mathbf{x}_{t_k}; t_k, e, y) - \hat{\mathbf{x}}_{\{0,e\}}. \tag{29}$$

It is straightforward that at the boundary timestep e, the error is,

$$e_e^e = \hat{\mathbf{x}}_{\{\pi,e\}} - \hat{\mathbf{x}}_{\{0,e\}}. \tag{30}$$

We then derive $e_e^{t_k}$ at t_k as:

$$\begin{aligned}
\boldsymbol{e}_e^{t_k} =& G_\theta\left(\mathbf{x}_{t_k}; t_k, e, y\right) - \hat{\mathbf{x}}_{\{0,e\}}, \\
=& G_\theta\left(\bar{\mathbf{x}}_{\{t_k \to t_{k+1}\}}; t_{k+1}, e, y\right) - \hat{\mathbf{x}}_{\{0,e\}}, \\
=& G_\theta\left(\bar{\mathbf{x}}_{\{t_k \to t_{k+1}\}}; t_{k+1}, e, y\right) - G_\theta\left(\mathbf{x}_{t_{k+1}}; t_{k+1}, e, y\right) + G_\theta\left(\mathbf{x}_{t_{k+1}}; t_{k+1}, e, y\right) - \hat{\mathbf{x}}_{\{0,e\}}, \\
=& G_\theta\left(\bar{\mathbf{x}}_{\{t_k \to t_{k+1}\}}; t_{k+1}, e, y\right) - G_\theta\left(\mathbf{x}_{t_{k+1}}; t_{k+1}, e, y\right) + \boldsymbol{e}_e^{t_{k+1}}.
\end{aligned} \tag{31}$$

According to Lipschitz condition on G_θ, we derive:

$$\begin{aligned}
\left\|\boldsymbol{e}_e^{t_k}\right\| &\le \left\|G_\theta\left(\bar{\mathbf{x}}_{\{t_k \to t_{k+1}\}}; t_{k+1}, e, y\right) - G_\theta\left(\mathbf{x}_{t_{k+1}}; t_{k+1}, e, y\right)\right\| + \left\|\boldsymbol{e}_e^{t_{k+1}}\right\|, \\
&\le L\left\|\bar{\mathbf{x}}_{\{t_k \to t_{k+1}\}} - \mathbf{x}_{t_{k+1}}\right\| + \left\|\boldsymbol{e}_e^{t_{k+1}}\right\|, \\
&\overset{(i)}{=} \left\|\boldsymbol{e}_e^{t_{k+1}}\right\| + \mathcal{O}\left((t_k - t_{k+1})^{p+1}\right),
\end{aligned} \tag{32}$$

where (i) *hold according to the local error of Euler solver [4, 6]. Iteratively, we can obtain the upper bound of distillation error at* e*:*

$$\begin{aligned}\|\boldsymbol{e}_e^e\|_2 &\leqslant \sum_{k=n_e}^{N-1} \mathcal{O}\left((t_{k+1}-t_k)^{p+1}\right) + \|e_e^T\|, \\ &\overset{(ii)}{\approx} \sum_{k=n_e}^{N-1} (t_{k+1}-t_k)\,\mathcal{O}\left((\Delta t)^p\right), \\ &= \mathcal{O}\left((\Delta t)^p\right) \sum_{k=n_e}^{N-1} (t_{k+1}-t_k), \\ &= \mathcal{O}\left((\Delta t)^p\right)(T-e),\end{aligned} \tag{33}$$

where N *is the index of time* T*,* (ii) *holds because* $\|e_e^T\|$ *describes the KL divergence between* $\mathbf{x}_0$ *and* $G_\theta\left(\hat{\mathbf{x}}_{\{\pi,T\}}; T, e, y\right)$*, where* $\hat{\mathbf{x}}_{\{\pi,T\}}$ *follows the normal distribution. For a well-trained diffusion model,* $\|e_e^T\|$ *should be bounded and remain constant at* e*, which can be ignored. The proof is completed.*

According to Lemma. 1, we prove that the proposed CC loss will converge within a lower error bound than CDS loss in [6] when $e > 0$.

11 Additional Ablation Studies

Effect of CC and CG. We conduct additional ablation studies to evaluate the effect of $\mathcal{L}_{\text{CC}}$ and $\mathcal{L}_{\text{CG}}$. Specifically, we compare $\mathcal{L}_{\text{CC}} + \mathcal{L}_{\text{CG}}$ with $\mathcal{L}_{\text{ISD}}$ (LucidDreamer [2]), $\mathcal{L}_{\text{CG}}$, and $\mathcal{L}_{\text{CG}} + \mathcal{L}^0_{\text{CC}}$, where we note $\mathcal{L}^0_{\text{CC}}$ as $\mathcal{L}_{\text{CC}}$ with $e \equiv 0$. As shown in Fig. 11, we find the proposed $\mathcal{L}_{\text{CG}}$ has higher generation quality compared to the LucidDreamer. $\mathcal{L}_{\text{CC}}$ will further add more details to the generated results in the generated 3D asset. Comparing with $\mathcal{L}_{\text{CG}} + \mathcal{L}^0_{\text{CC}}$ that shares the same distillation error bound with CDS [6], we found the color distortion becomes more severe and tend to be over-exposed, validating the effect of proposed $\mathcal{L}_{\text{CC}}$ in reducing distillation error.

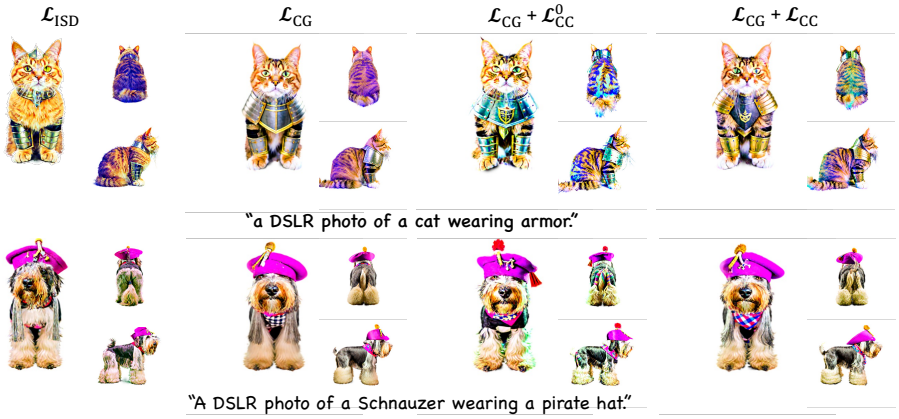

Fig. 11: Qualitative comparison regarding different variances of $\mathcal{L}_{\text{CC}}$ and $\mathcal{L}_{\text{CG}}$. From left to right, 3D asset generated by $\mathcal{L}_{\text{ISD}}$, $\mathcal{L}_{\text{CG}}$, $\mathcal{L}_{\text{CG}} + \mathcal{L}^0_{\text{CC}}$, and $\mathcal{L}_{\text{CG}} + \mathcal{L}_{\text{CC}}$, respectively.

Effect of High Order ODE-Solver. We try 2nd order DPM-Solver at denoising steps, and the results indicate differences in illumination and details, as shown in Fig. 12.

Fig. 12: Generated views of GCS+BEG under a low CFG weight ($w = 7.5$) by using first-order (right) and second-order DPM-Solver.

12 Additional Qualitative Evaluation

In this section, we show more quality comparison between the proposed GCS, DreamFusion [3], GaussianDreamer [7], ProlificDreamer [5], and LucidDreamer [2] in Fig. 13.

Fig. 13: Additional qualitative comparison among DreamFusion [3] (column 1), GaussianDreamer [7] (column 2), ProlificDreamer [5] (column 3), LucidDreamer [2] (column 4), and our method (column 5).

References

1. Kim, D., Lai, C.H., Liao, W.H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., Ermon, S.: Consistency trajectory models: Learning probability flow ode trajectory of diffusion. Int. Conf. Learn. Represent. (2024)
2. Liang, Y., Yang, X., Lin, J., Li, H., Xu, X., Chen, Y.: Luciddreamer: Towards high-fidelity text-to-3D generation via interval score matching. arXiv preprint arXiv:2311.11284 (2023)
3. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3D using 2d diffusion. In: Int. Conf. Learn. Represent. (2022)
4. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. arXiv preprint arXiv:2303.01469 (2023)
5. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. Adv. Neural Inform. Process. Syst. (2024)
6. Wu, Z., Zhou, P., Yi, X., Yuan, X., Zhang, H.: Consistent3D: Towards consistent high-fidelity text-to-3D generation with deterministic sampling prior. arXiv preprint arXiv:2401.09050 (2024)
7. Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., Wang, X.: Gaussiandreamer: Fast generation from text to 3D gaussian splatting with point cloud priors. IEEE Conf. Comput. Vis. Pattern Recog. (2024)
8. Zheng, J., Hu, M., Fan, Z., Wang, C., Ding, C., Tao, D., Cham, T.J.: Trajectory consistency distillation. arXiv preprint arXiv:2402.19159 (2024)