4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling

Sherwin Bahmani[1,2] Ivan Skorokhodov[3,4] Victor Rong[1,2] Gordon Wetzstein[5] Leonidas Guibas[5]
Peter Wonka[3] Sergey Tulyakov[4] Jeong Joon Park[6] Andrea Tagliasacchi[1,7,8] David B. Lindell[1,2]
[1]University of Toronto [2]Vector Institute [3]KAUST [4]Snap Inc. [5]Stanford University [6]University of Michigan [7]SFU [8]Google

1. Implementation Details

Here, we provide additional implementation details.

Optimization Following [4, 6], during the first stage of the optimization, we anneal the sampled diffusion timesteps over 5000 iterations from $t_d \in [0.02, 0.98]$ to a range of $t_d \in [0.02, 0.5]$. In subsequent stages, we keep the timestep sampling the same, except for the video SDS updates, where we sample $t_d \in [0.02, 0.98]$. For the $\nabla_\theta \mathcal{L}_{\text{VID}}$ updates, we set the total learning rate to 0.1, while $\nabla_\theta \mathcal{L}_{\text{3D}}$ and $\nabla_\theta \mathcal{L}_{\text{IMG}}$ use a learning rate of 1.0. To render a background for the image, we optimize a second small MLP that takes in a ray direction and returns a background color. We composite the ray color $\mathbf{C}$ rendered from $\mathcal{N}_\theta$ on top of this background color.

Model Our hash-grid-based representation [3] uses default values [1] for both the static and dynamic components: 16 levels, 2 features per level, and a base resolution of 16.

Rendering We use NerfAcc [2] as our rendering pipeline without any changes to the default values in threestudio [1]. During training and inference we use 512 samples per ray. Following the MAV3D [5] evaluation protocol, we sample 64 frames during inference after training our model with 16 frames, which is possible due to the implicit time coordinate. During training, we sample a random time offset with evenly spaced time coordinates over the [0.0, 1.0] time interval. This enables smooth time sampling across the whole range of time coordinates.

Computation. We optimized the model on an NVIDIA A100 GPU. The entire procedure requires roughly 80 GB of VRAM and the three optimization stages require approximately 2, 2, and 19 hours of compute, respectively.

Runtime. Rendering one frame for our method takes 71ms vs. 68 ms for a static 3D MVDream based model. We cannot compare to MAV3D as code is not available, and they do not report runtimes.

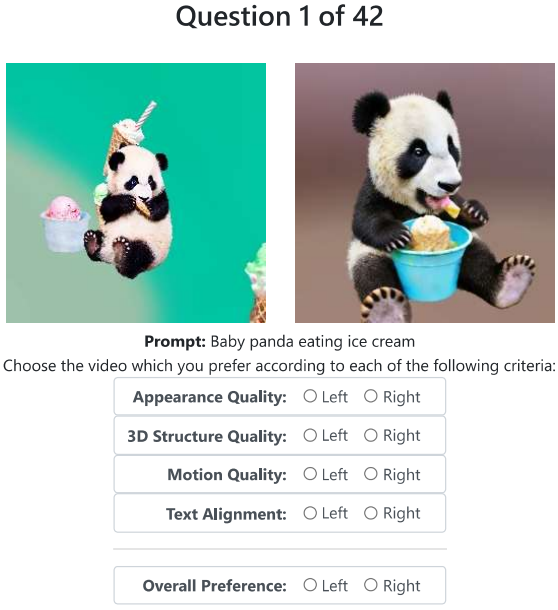

Figure 1. **Example survey question.** A survey question is shown above. The two videos are rendered using a baseline method (left) and 4D-fy (right). The left–right ordering of videos generated using each method is randomized throughout the survey.

2. User Study

The user study was carried out as a single survey consisting of 53 questions, each with 5 subquestions. Each question asked the evaluator to compare two videos: one showing a scene rendered with 4D-fy and another rendered with a separate method using the same text prompt as input. Evaluators filled out five subquestions that asked their preferred video based on appearance quality, 3D structure quality, motion quality, text alignment, and overall preference, as shown in Fig. 1. Evaluators were given the following instructions for each metric.

- **Appearance Quality:** Evaluate the clarity and visual appeal of the scene as it appears from any particular viewpoint (ignoring, e.g., inconsistencies in appearance across

Table 1. Further ablations: We indicate % users who prefer 4D-fy over each method (Col. 1).

Method	*CLIP*	AQ	SQ	MQ	TA	Overall
4D-fy	35.0		—			—
w/ single-stage	29.7	100%	100%	88%	97%	100%
w/ single hash grid	25.5	100%	100%	99%	100%	100%

different viewpoints). Your assessment should focus on the appearance of the foreground object and ignore the background of the video.

- **3D Structure Quality:** Assess the detail and realism of the shape of the scene across the multiple viewpoints shown in the video.
- **Motion Quality:** Assess the realism of motion, including the amount of motion and how naturally the movements in the video are portrayed.
- **Text Alignment:** Determine how accurately each video reflects the content of the text prompt. Consider whether the key elements of the prompt are represented.
- **Overall Preference:** State your overall preference between the two videos. This is your subjective appraisal of which video, in your view, stands out as better based on appearance quality, 3D structure quality, motion quality, and text alignment, (i.e., overall quality).

Of the 53 questions, 28 were comparisons between 4D-fy and MAV3D. The other 25 questions were between 4D-fy and each of the five ablated methods described in Sec. 4.3 of the main paper, with five questions for each. For each comparison and metric, the results were tested against the null hypothesis that evaluators had no preference between either method; i.e., they would choose either with probability 0.5. We aggregate over prompts and evaluators and use χ^2 analysis to determine the corresponding p-value. We choose $p < 0.05$ as a significant deviation from the null hypothesis and find that—with the exception of comparing motion quality between 4D-fy and MAV3D—all p-values are well below 0.05. All statistically significant results indicated that users preferred videos rendered using 4D-fy over MAV3D.

3. Qualitative Results

In Fig. 2, we present additional results generated using our method. We highly encourage readers to view the videos included in our supplementary website to gain a better appreciation of our text-to-4D generation results.

4. Ablations

In Tab. 4 and Fig. 3 we show further ablation results. Training naively only the last stage using all three diffusion models leads to poor quality, as the model struggles to simultaneously learn 3D structure, appearance, and motion. Using a large single dynamic hash grid instead of decomposed smaller static and dynamic hash grids also leads to significantly lower quality results.

5. Geometry

In Fig. 4, we show normals and meshes extracted with marching cubes. While generating high-quality geometry is not the main goal of this work, our method shows comparable quality to previous text-to-3D methods.

6. Limitations

In addition to limitations outlined in the main paper, we briefly discuss the temporal flickering in renderings generated with 4D-fy. The main goal of our work is to generate high-quality dynamic 3D scenes—and especially to improve image quality compared to previous work where results can appear overly smooth or cartoon-like (e.g., Fig. 4). It may be possible to mitigate flickering artifacts by placing more emphasis on supervision with the video diffusion model; however, this may trade off image quality. Reducing temporal flickering without any penalty to image quality (e.g., avoiding blurry or overly smooth images) is an important direction for future work.

7. Ethics Statement

In its current form, our method is not able to edit real people. However, it could be extended and misused for generating edited imagery of real people. We condemn the application of our method for creating realistic fake content intended to harm specific entities or propagate misinformation.

References

[1] Threestudio Github page. https://github.com/threestudio-project/threestudio. Accessed: 2023-10-31. 1

[2] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *Proc. ICCV*, 2023. 1

[3] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 2022. 1

[4] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 1

[5] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. In *Proc. ICML*, 2023. 1

[6] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. *Proc. NeurIPS*, 2023. 1

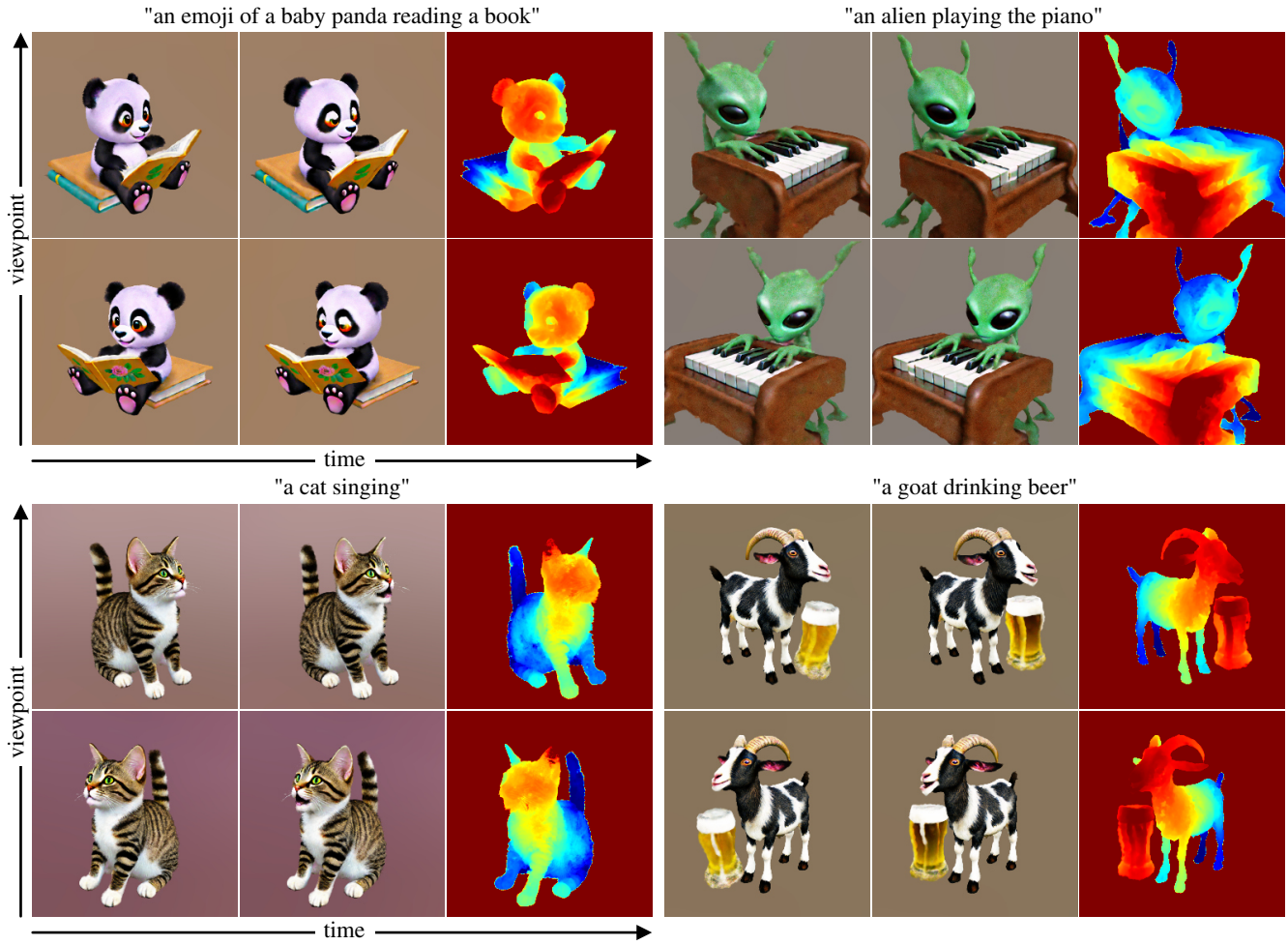

Figure 2. **Text-to-4D Synthesis.** We present additional results using 4D-fy.

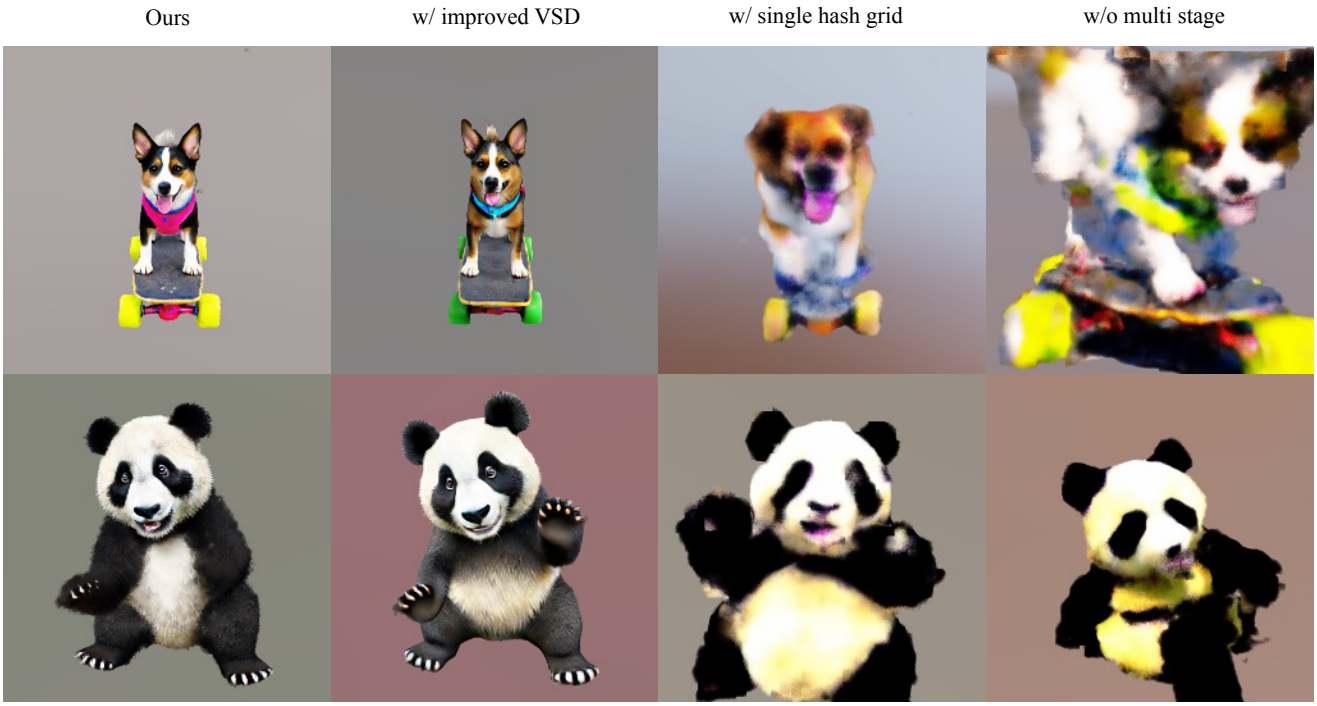

Figure 3. Further ablation studies (zoom in for details). We also show that a slightly higher learning rate for VSD can improve the visual details.

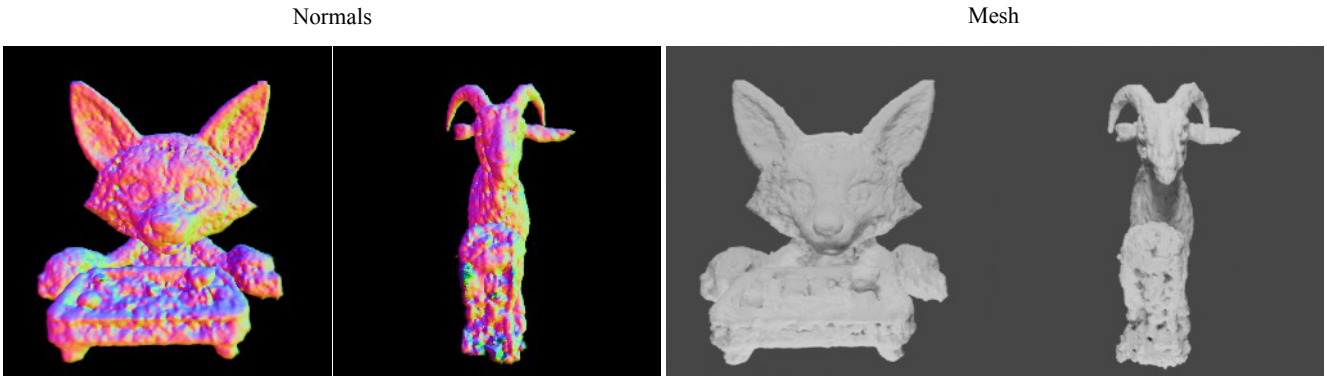

Figure 4. Normals and meshes.