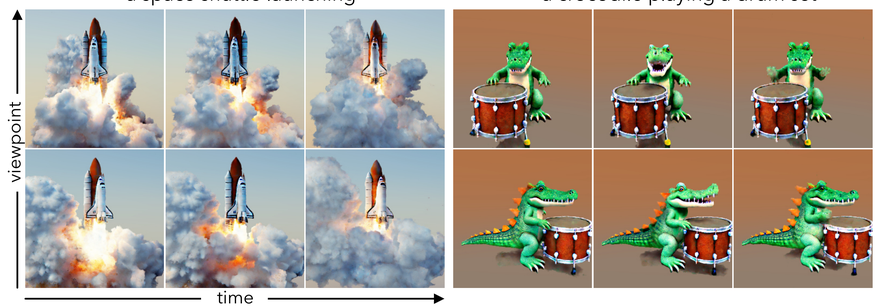
4D-fy: Text-to-4D Generation Using Hybrid Score Distillation Sampling

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

Website

Motivation

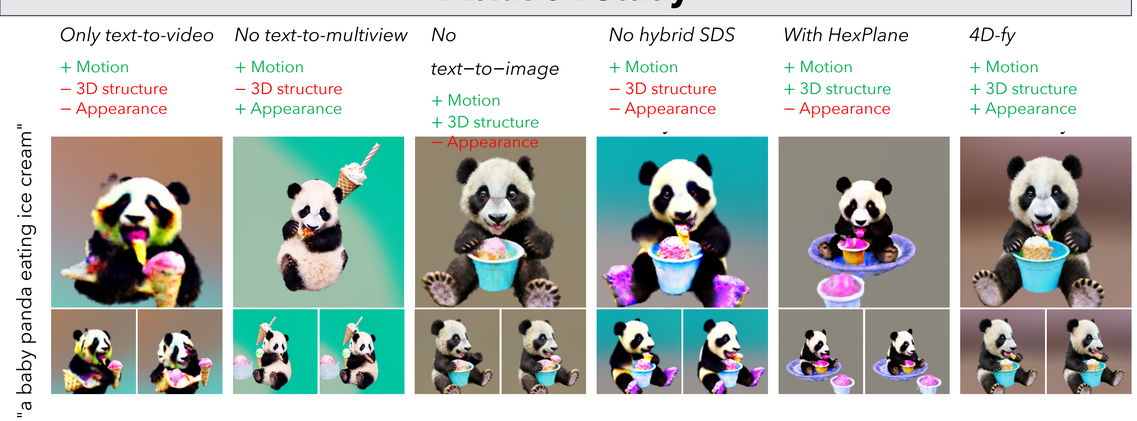


We generate text-to-4D scenes using alternating optimization (hybrid SDS) between pre-trained (i) text-to-video, (ii) text-to-multiview, and (iii) text-to-image models.

Challenges:

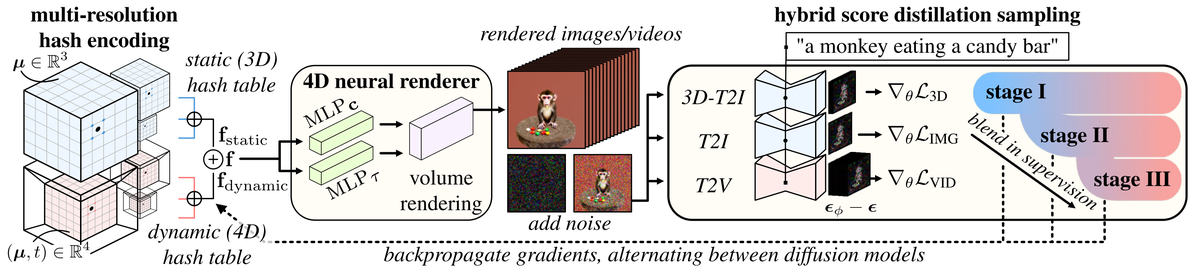
- **Motion**. The scene is animated using a text-to-video model (ZeroScope).
- **3D structure**. The text-to-multiview model addresses the multi-face problem (MVDream).
- **Appearance**. Single-view text-to-image models offer the best visual quality (StableDiffusion).

Ablation Study



Method

4D-fy uses a three-stage training process that smoothly blends supervision from different pre-trained diffusion models using alternating optimization.



Text-to-4D Generation