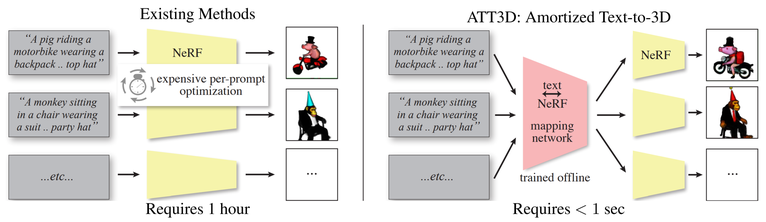
ATT3D: Amortized Text-to-3D Object Synthesis

Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa
Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, James Lucas

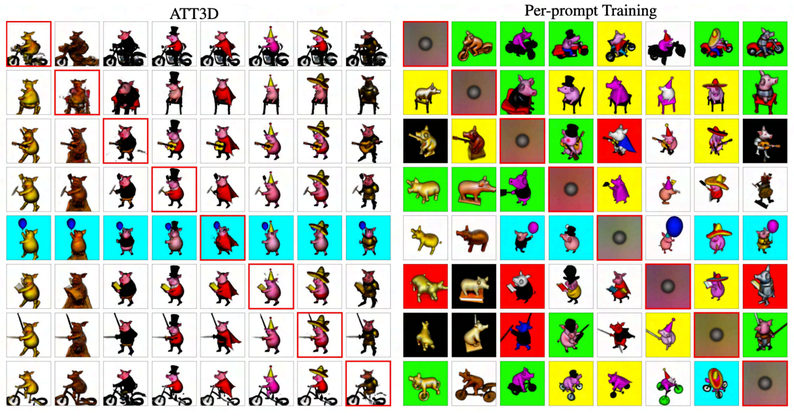
Overview

- Recent work[1–5] generates high-quality 3D objects from text.
- But these require a slow per-prompt optimization.
- We generate from unseen prompts fast via an "amortized" model trained on many prompts simultaneously.

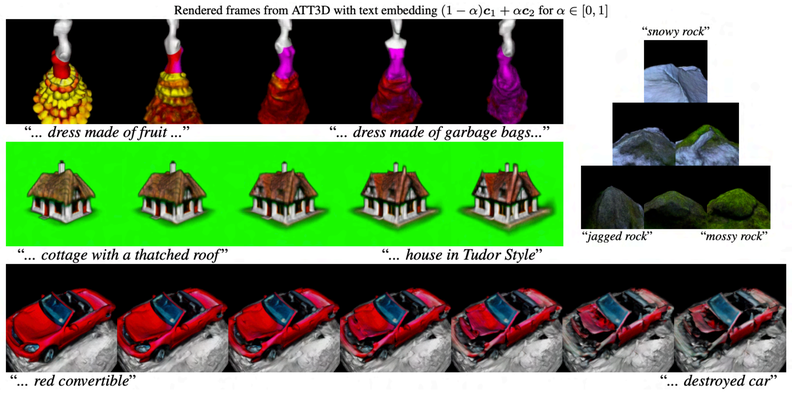


Benefits

- **Fast**: We generalize to - i.e., no optimization on - unseen testing prompts along the diagonal in red for a compositional dataset with template "a pig {activity} {theme}."

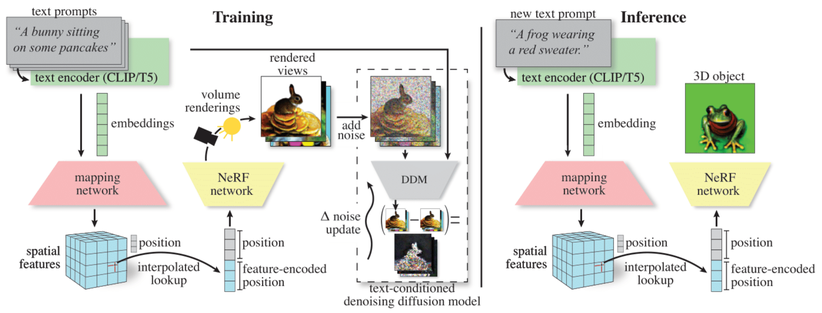


- **Interpolations**: We create 3D animations and continuums of novel assets by interpolating between prompt embeddings.



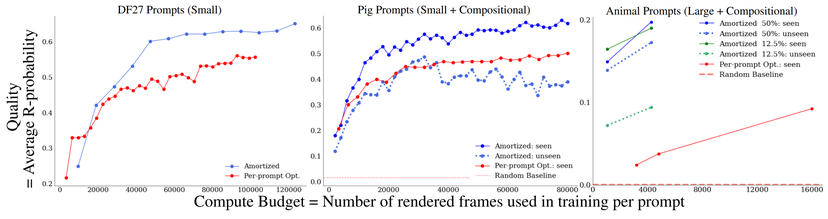
Our Method: ATT3D

- We augment the text-to-3D pipeline to re-use the text-to-image model's text-embedding to condition our 3D representation.
- We use a NeRF with an instant-NGP backbone, whose spatial weights we output via a hypernetwork using text-embeddings.

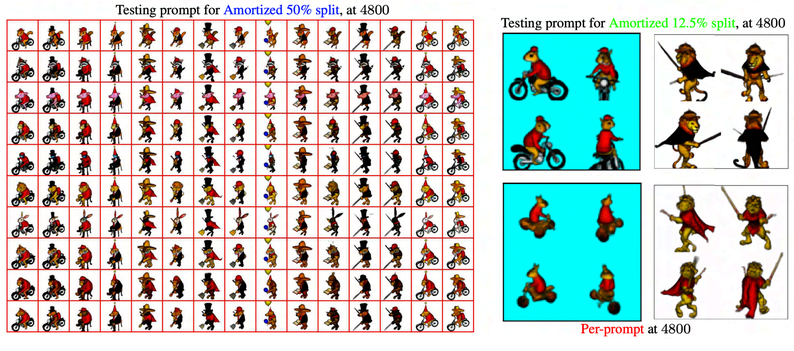


Measuring Memorization and Generalization

- For any compute budget, we achieve a higher quality on both the seen and unseen prompts, with growing benefits for larger, compositional prompt sets.



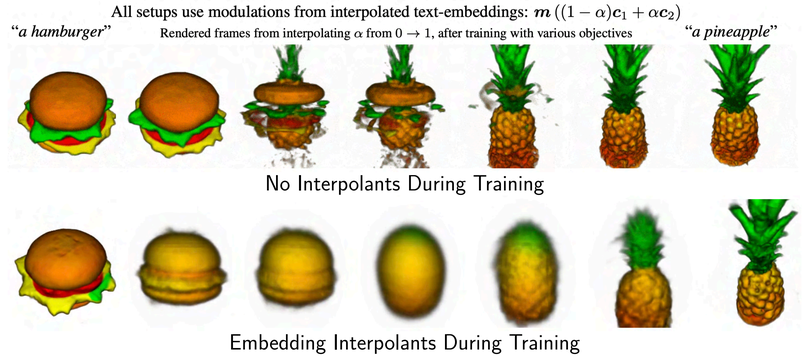
- We show animals for prompt combos unseen during training.



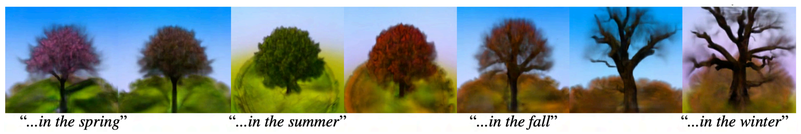
- We generate 3D object + render from unseen text in real-time.
- Our method might be more robust to underfitting (failed prompts) and some types of overfitting (Janus faces).

Interpolations

- To improve interpolant quality, we sample them during training.
- Various interpolant strategies – ex., loss, embedding, or guidance.



- We interpolate on lists of prompts for more complex results.



- Animations and more examples on project website in QR code.

Future Directions

- Boost quality and get diverse output on functioning prompts.
  - Combining with concurrent works can help solve these.
- Make process robustly work for different prompts (underfitting).
- Fix Janus face artifacts in output (overfitting).
- Improve interpolations to be semantically meaningful for effective text-to-4D generation.

References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv:2303.13873*, 2023.

[2] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022.

[4] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] 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. *arXiv:2305.16213*, 2023.