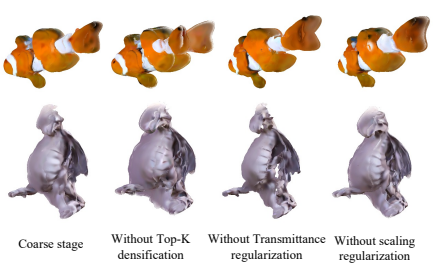
Vista3D: Unravel the 3D Darkside of a Single Image

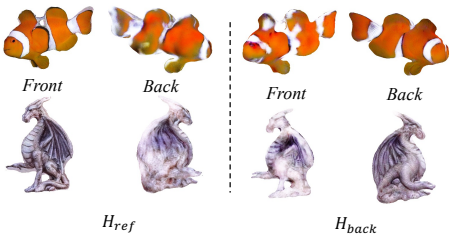
Supplementary Material

1 More experimental results

1.1 More ablation studies



(a) Ablation study of the coarse stage. Here we conduct four settings on the coarse stage, including w/o Top-K densification, w/o transmittance and scaling regularization for comparison.



(b) Visualization of the disentangled texture. Here we showcase a generated 3D object. The left side is visualized from the facing-forward hash encoding H_{ref}, while the right side is visualized from the back hash encoding H_{back}.

Fig. 6: Ablation study of the coarse stage and disentangled texture.

Top-k densification. We compare our densification strategy against a naive gradient threshold approach. This comparison is illustrated in the second column of Figure 6a. Using a naive gradient threshold often results in excessive densification of 3D Gaussians, causing geometry to appear swollen. Furthermore, finding an appropriate gradient threshold is challenging, as it varies from case to case. In contrast, our method deterministically controls the densification ratio throughout the optimization process. Consequently, the total number of 3D Gaussians at convergence is solely influenced by the hyperparameter of pruning opacity, effectively maintaining the number of 3D Gaussians within a reasonable range and yielding more accurate geometry.

Regularization with 3DGS. In the third and fourth columns of Figure 6a, we conduct ablation experiments on the two regularization terms specified in Equation 2: transmittance regularization and scale regularization. Removing the transmittance regularization tends to produce objects with holes, resulting in coarse meshes from these 3D Gaussians that are often not watertight, complicating refinement stage optimization. On the other hand, excluding only the scale regularization often leads to coarser details in the geometry. This may be caused by Gaussians with larger scales oversmoothing the local geometries.

The effect of prior composition. To explore the 3D dark side of a single image, we introduce a gradient constraint-based method in Sec. 3.3 to control two diffusion priors in the image-to-3D task. Here we conduct an ablation study to validate the effectiveness of this component. As shown in Fig. 7, without this

score composition, though detailed texture on the backside can still be generated, results in degraded consistency between front views and reference images. Another setting involves a naive weighting strategy; we follow Magic123 [28] to set a weighting factor of 1/40 on the SDS term $\mathcal{L}^{\rho}_{SDS}$ with diffusion prior ϵ_{ρ}. With this setting, the backside of the generated 3D objects appears overly smoothed. In contrast, incorporating score composition enables our Vista3D to robustly generate textures that are both detailed and consistent across the front and back views of 3D objects.

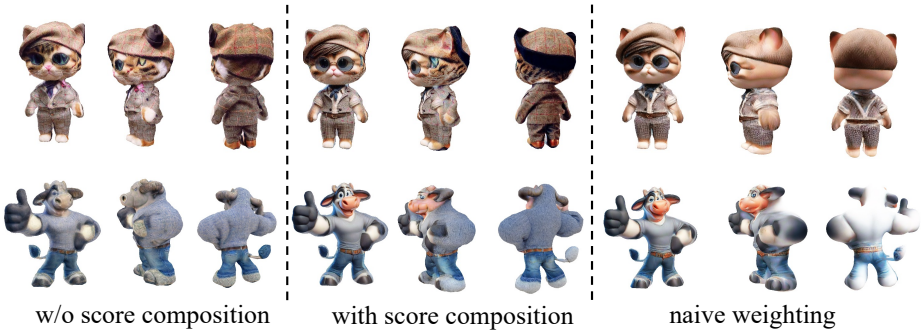


Fig. 7: Ablation Study of Score Composition. Without score composition, the consistency between the reference view and front view is degraded. Applying naive weighting results in over-smoothed textures on back views.

1.2 More qualitative results

Figure 8 showcases the qualitative results of Vista3D-L with diffusion prior composition compared to Vista3D-S with a single diffusion prior. Particularly in scenarios where the provided reference view is less informative, such as when only a side or back view of an object is available, Vista3D-L demonstrates a superior ability to generate more detailed textures compared to Vista3D-S, especially when specific text prompts are used. For example, in the case of the astronaut, Vista3D-S tends to produce oversmoothed textures. In contrast, when using Vista3D-L, the textures generated are notably more vivid and detailed.

2 Camera Pose Sampling

As illustrated in Fig. 9, our approach adopts a 3D-aware camera pose sampling strategy in the refinement stage, diverging from the standard uniform sampling used in previous image-to-3D works [28, 38, 39]. This approach not only speeds up convergence but also enhances visual quality.

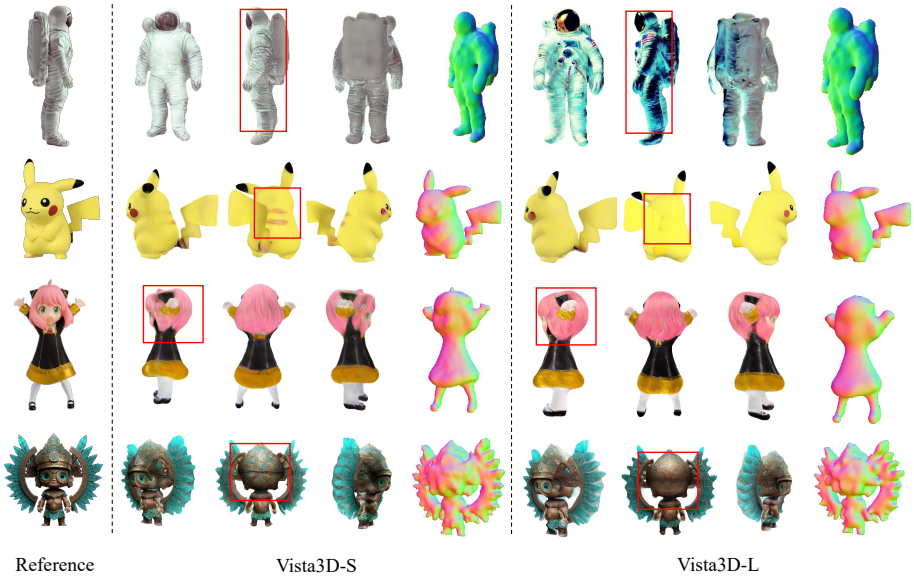


Fig. 8: Qualitative Comparison between Vista3D-S and Vista3D-L

Specifically, for a given conditional reference image I_{ref}, the pre-trained Zero-1-to-3 model [18] ϵ_ϕ is capable of approximating the underlying 3D object distribution $P_{I_{ref}}(x)$. Leveraging this, we employ its estimated empirical error for 3D-aware sampling.

In this sampling stage, camera poses are sampled from a sphere surface surrounding the central object, divided evenly into N sub-regions R_i with azimuth ranging from $[-180, 180]$ degrees, as shown on the left side of Figure 9. Memory queues of fixed length T are established for each sub-region to store empirical errors estimated during the SDS optimization, directly derived from SDS as $(\epsilon_\phi - \epsilon)$ in Equation 1.

When performing pose sampling, an empirical Probability Density Function (PDF) $P_{3d}(R_i)$ is created from these N memory queues. Additionally, given the supplementary supervision from the reference image I_{ref} for forward-facing camera poses, we integrate Gaussian unsampling to reduce sampling frequency on forward-facing poses and increase it for unseen views. This unsampling employs a rejection sampling with a truncated Gaussian distribution, depicted on the right side of Figure 9. Each sub-region is mapped onto this truncated Gaussian PDF, with regions overlapping significantly with the reference view being more likely to be sampled.

In this process, a camera pose is sampled by initially performing Gaussian unsampling to determine a rejection index $n \in [0, N-1]$. Subsequently, we modify the empirical PDF by setting $P_{3d}(R_n) = 0$ and normalizing it. A sub-region index is then sampled from this discrete PDF $\tilde{P}_{3d}(R_i)$, and a camera pose is uniformly sampled from this chosen sub-region.

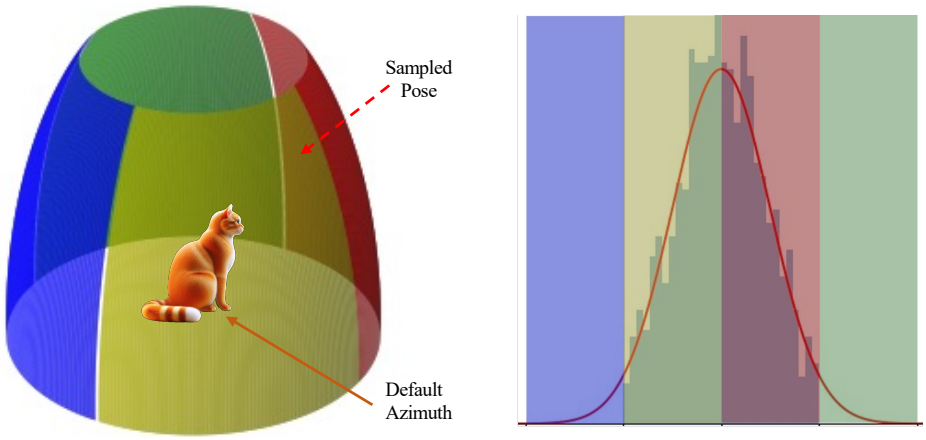


Fig. 9: 3D-aware Pose Sampling, Camera poses are sampled from an empirical PDF with a truncated Gaussian unsampling.

In our implementation, we configure $N = 5$, and initially perform uniform camera pose sampling during the first 100 iterations. For the Gaussian Unsampling, we utilize a truncated Gaussian distribution spanning $[-1, 1]$, with $\mathcal{N}(0, 0.5)$. This distribution is evenly divided into N intervals to facilitate the sampling process.

3 Timestep Sampling in SDS

Pioneering work DreamFusion [27] randomly sample timestep t from $\mathcal{U}(20, 980)$ in the SDS optimization. However, Dreamtime [10] critiques this strategy, suggesting that such random sampling is misaligned with the Denoising Diffusion Probabilistic Models (DDPM) sampling process and leads to inefficient and inaccurate optimization in SDS. Dreamtime suggests a deterministic Time Prioritized (TP) strategy where each iteration step is assigned a unique, decrementally decreasing timestep t.

However, we observed that this deterministic approach falls short in SDS optimization. Artifacts generated by large timesteps are not effectively compensated for by smaller timesteps, often exacerbating the problem. To rectify this, we propose an interval-based annealing method for the timestep. Specifically, we define a maximum timestep t_{max} and a minimum timestep t_{min} for each optimization interval, updating them every 50 optimization steps. The timestep is then sampled from the dynamically adjusted interval $\mathcal{U}(t_{min}, t_{max})$. This approach effectively alleviates the artifacts that larger timesteps tend to cause.

4 Limitations

Despite Vista3D demonstrating prowess in exploring the 3D dark side of a single image, we acknowledge several limitations for future exploration. Employing a Score Distillation Sampling (SDS) based architecture, Vista3D necessitates optimization for each 3D object it generates, positioning its efficiency a notch below that of purely feed-forward image-to-3D methods. The amount of public 3D data is relatively limited, often resulting in the generation of simplistic 3D objects by feed-forward methodologies. Vista3D leverages diffusion prior composition to facilitate the reconstruction of more diverse 3D objects. This strategy holds promise for the creation of additional 3D data, potentially alleviating the current data scarcity and enabling the development of more sophisticated pre-trained image-to-3D models.

References

1. Cao, Z., Hong, F., Wu, T., Pan, L., Liu, Z.: Large-vocabulary 3d diffusion model with transformer. arXiv preprint arXiv:2309.07920 (2023)
2. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In: ICCV (October 2023)
3. Chen, Z., Wang, F., Liu, H.: Text-to-3d using gaussian splatting. arXiv preprint arXiv:2309.16585 (2023)
4. Chung, J., Lee, S., Nam, H., Lee, J., Lee, K.M.: Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384 (2023)
5. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)
6. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR. pp. 13142–13153 (2023)
7. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022)
8. Duggal, S., Pathak, D.: Topologically-aware deformation fields for single-view 3d reconstruction. In: CVPR. pp. 1536–1546 (2022)
9. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
10. Huang, Y., Wang, J., Shi, Y., Qi, X., Zha, Z.J., Zhang, L.: Dreamtime: An improved optimization strategy for text-to-3d content creation. arXiv preprint arXiv:2306.12422 (2023)
11. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: CVPR. pp. 857–866. IEEE (2022)
12. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
13. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
15. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR. pp. 300–309 (2023)
16. Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion (2023)
17. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., et al.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928 (2023)
18. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)

19. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
20. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023)
21. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp. 347–353 (1998)
22. Melas-Kyriazi, L., Laina, I., Rupprecht, C., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image. In: CVPR. pp. 8446–8455 (2023)
23. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
24. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG) **41**(4), 1–15 (2022)
25. Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., Chen, M.: Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022)
26. Nielson, G.M.: Dual marching cubes. In: IEEE visualization 2004. pp. 489–496. IEEE (2004)
27. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: ICLR. OpenReview.net (2023)
28. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843 (2023)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
30. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., et al.: Zeronvs: Zero-shot 360-degree view synthesis from a single real image. arXiv preprint arXiv:2310.17994 (2023)
31. Seo, J., Jang, W., Kwak, M.S., Ko, J., Kim, H., Kim, J., Kim, J.H., Lee, J., Kim, S.: Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. arXiv preprint arXiv:2303.07937 (2023)
32. Shen, Q., Yang, X., Wang, X.: Anything-3d: Towards single-view anything reconstruction in the wild (2023)
33. Shen, Q., Yi, X., Wu, Z., Zhou, P., Zhang, H., Yan, S., Wang, X.: Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. arXiv preprint arXiv:2403.18795 (2024)
34. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems **34**, 6087–6101 (2021)
35. Shen, T., Munkberg, J., Hasselgren, J., Yin, K., Wang, Z., Chen, W., Gojcic, Z., Fidler, S., Sharp, N., Gao, J.: Flexible isosurface extraction for gradient-based mesh optimization. ACM Trans. Graph. **42**(4), 37–1 (2023)
36. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)

37. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. arXiv preprint arXiv:2310.16818 (2023)
38. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
39. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In: ICCV. pp. 22819–22829 (October 2023)
40. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: CVPR. pp. 12619–12629 (2023)
41. Wu, Z., Zhou, P., Yi, X., Yuan, X., Zhang, H.: Consistent3d: Towards consistent high-fidelity text-to-3d generation with deterministic sampling prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9892–9902 (2024)
42. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024)
43. Xu, Y., Tan, H., Luan, F., Bi, S., Wang, P., Li, J., Shi, Z., Sunkavalli, K., Wetzstein, G., Xu, Z., et al.: Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. arXiv preprint arXiv:2311.09217 (2023)
44. Yang, X., Wang, X.: Hash3d: Training-free acceleration for 3d generation. arXiv preprint arXiv:2404.06091 (2024)
45. 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. arXiv preprint arXiv:2310.08529 (2023)
46. Yi, X., Wu, Z., Shen, Q., Xu, Q., Zhou, P., Lim, J.H., Yan, S., Wang, X., Zhang, H.: Mvgamba: Unify 3d content generation as state space sequence modeling. arXiv preprint arXiv:2406.06367 (2024)
47. Yi, X., Wu, Z., Xu, Q., Zhou, P., Lim, J.H., Zhang, H.: Diffusion time-step curriculum for one image to 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9948–9958 (2024)
48. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR. pp. 4578–4587 (2021)
49. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
50. Zou, Z.X., Yu, Z., Guo, Y.C., Li, Y., Liang, D., Cao, Y.P., Zhang, S.H.: Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10324–10335 (2024)