Supplementary Material of ThemeStation: Generating Theme-Aware 3D Assets from Few Exemplars

ZHENWEI WANG, City University of Hong Kong, China
TENGFEI WANG*, Shanghai Artifcial Intelligence Laboratory, China
GERHARD HANCKE, City University of Hong Kong, China
ZIWEI LIU, S-Lab, Nanyang Technological University, Singapore
RYNSON W.H. LAU*, City University of Hong Kong, China

Code & video: https://3dthemestation.github.io/

A IMPLEMENTATION DETAILS

In the first stage, we render 20 images for each reference model with a fixed elevation, *i.e.,* 0 or 20, and randomized azimuth. We fine-tune the pre-trained Stable Diffusion [Rombach et al. 2022] model for 200 iterations (a single exemplar) or 400 iterations (a few exemplars) with a batch size of 8. We set the learning rate as 2×10^{-6}, the image size as 512×512, and the CFG weight at inference as 7.5. We also take the camera pose of the rendered images as an additional condition during the model fine-tuning step to ensure the generated concept images have a correct viewpoint for accurate image-to-3D initialization.

In the second stage, we employ an off-the-shelf image-to-3D method [Long et al. 2023] to lift the synthesized concept image into an initial 3D model, represented as a neural implicit signed distance field (SDF). We use the concept image and 20 augmented views of the initial model for concept prior learning and use 30 normal maps, and 30 color images of the input 3D exemplars for reference prior learning. During optimization, we convert the SDF into DMTet [Shen et al. 2021] at a 192 grid and 512 resolution to directly optimize the textured mesh at each optimization iteration. We render both the normal map and the color image, under randomized viewpoints, as guidance to compute the DSD loss (Eq. 5). We use dynamic diffusion timestep that samples larger timestep from range $[0.5, 0.75]$ when applying the concept prior and samples smaller timestep from range $[0.1, 0.25]$ for the reference prior. We set α as 0.2 and β as 1.0. The total optimization step is 5000. We also adopt the total variation loss [Rudin et al. 1992] and contextual loss [Mechrez et al. 2018] to enhance the texture quality. Specially, the contextual loss is applied between the rendered color image and the 20 augmented views of the initial model. The whole 3D-to-3D generation process takes around 2 hours using a single NVIDIA A100 GPU.

B USER STUDY SETTINGS

We randomly select 20 models from our dataset and generate 3 variations for each model. We invite a total of 30 users, recruited publicly, to complete a questionnaire consisting of 30 pairwise comparisons (15 for image-to-3D and 15 for 3D-to-3D) in person, totaling 900 answers. For image-to-3D, we show two generated 3D models (one by our method and one by the baseline method) beside a concept image and ask the users to answer the question: "Which of the two models do you prefer (*e.g.,* higher quality and more details) on the premise of aligning with the input view?" For 3D-to-3D, we show two sets of generated 3D variations beside a reference model and ask the question: "Which of the two sets do you prefer (*e.g.,* higher quality and more diversity) on the premise of sharing consistent themes with the reference?"

Table 1. Quantitative evaluation of theme-driven diffusion model.

	Iteration100	Iteration200	Iteration300	Iteration400
LPIPS-diversity ↑	0.627	0.617	0.403	0.347
LAION-aesthetic-score ↑	6.262	6.355	6.367	5.941

C EVALUATION OF THEME-DRIVEN DIFFUSION MODEL

To evaluate the influence of different fine-tuning iterations for the theme-driven diffusion model that generates concept images in the first stage, we conduct ablation studies on four settings, *i.e.,* fine-tuning the theme-driven diffusion model given one 3D exemplar for 100, 200, 300 and 400 iterations. We use LPIPS-diversity (LPIPS differences across generated images) and LAION-aesthetic-score to estimate the diversity and quality of generated concept images under different settings. The quantitative results are shown in Tab. 1. As can be seen, diversity significantly drops when iteration is 300, and quality drops when iteration is 400, both caused by overfitting. We thus set the fine-tuning iteration to 200 for a single exemplar (Sec. A).

D POTENTIAL ETHICS ISSUES

As a generative model, *ThemeStation* may pose ethical issues if used to create baleful and fake content, which requires more vigilance and care. We can adopt the commonly used safety checker in existing text-to-image diffusion models to filter out maliciously generated concept images in our first stage to alleviate the potential ethics issues.

[*] Co-corresponding authors.

Authors' addresses: Zhenwei Wang, zhenwwang2-c@my.cityu.edu.hk, City University of Hong Kong, Hong Kong SAR, China; Tengfei Wang, tfwang@connect.ust.hk, Shanghai Artifcial Intelligence Laboratory, Shanghai, China; Gerhard Hancke, gp.hancke@cityu.edu.hk, City University of Hong Kong, Hong Kong SAR, China; Ziwei Liu, ziwei.liu@ntu.edu.sg, S-Lab, Nanyang Technological University, Singapore; Rynson W.H. Lau, rynson.lau@cityu.edu.hk, City University of Hong Kong, Hong Kong SAR, China.

REFERENCES

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2023. Wonder3D: Single image to 3D using cross-domain diffusion. *arXiv preprint arXiv:2310.15008* (2023).

Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*. 768–783.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena* 60, 1-4 (1992), 259–268.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3D shape synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 6087–6101.