Appendix of Conditional 3D shape Generation based on Shape-Image-Text Aligned Latent Representation

Anonymous Author(s)
Affiliation
Address
email

This appendix serves as a supplementary extension, enriching and expanding upon the core content presented in the main body. We first describe the training details of the shape-image-text aligned auto-encoder (SITA-VAE) and aligned shape latent diffusion model (ASLDM) in section A. In section B, we describe more details for the zero-shot classification experiments in Figure 5 in the main text. Furthermore, in section C, we provide the predefined phrases for augmenting the shape-image-text data pair. Benefiting from the alignment among 3D shapes, images, and texts via contrastive learning, our model can retrieve 3D shapes given a query image, and we show the visual result in section D. We also show more visual comparisons in section E. Moreover, we test our model with conditioning input from the internet and show results in section F. Note that HTML files in the zip file accompany all visual results in browsers with interactive 3D viewing.

A Training Details

Stage 1: SITA-VAE. The encoder takes $N = 4096$ point clouds with normal features as the inputs. Equation (3) is the training loss for SITA-VAE. We set λ_c as 0.1 and λ_{KL} as 0.001. For the reconstruction term L_r, we follow the training strategies with 3DILG [7], which first normalize all mesh into $[-1, 1]$, and then separately samples 1024 volumetric points and 1024 near-surface points with ground-truth inside/outside labels from the watertight mesh. The mini-batch size is 40, and we train this model around 200,000 steps.

Stage 2: ASLDM. We the training diffusion scheduler with LDM [4] whose training diffusion steps are 1000, $\beta \in [0.00085, 0.012]$ with scaled linear β scheduler. The mini-batch size is 64, and we train the model around 500,000 steps. In the inference phase, we follow with the classifier-free guidance (CFG) [3] as shown in Equation (5), and we set the guidance scale λ as 7.5.

B Details in zero-shot classification experiments

Dataset. We conduct zero-shot classification experiments on ModelNet40 [5], which provides 12311 synthetic 3D CAD models in 40 categories. The dataset splits into two parts for training and testing, respectively, where the training set contains 9843 models and the testing set contains 2468 models.

Settings. We first train our shape-image-text aligned variational auto-encoder (SITA-VAE) on shapent [1]. Then, we utilize the trained encoders of SITA-VAE for classification on the testing set of ModelNet40 directly. Specifically, for a query 3D shape, we compute the cosine similarity between the shape and each category, where the category reformulates by the phrase "a 3D model of {}". Besides, we report top-1 accuracy and top-5 accuracy, where top-1 accuracy indicates that the ground-truth category achieves the highest similarity, and top-5 accuracy indicates that the ground-truth category achieves similarity in the top 5.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

C Template in building shape-image-text data pair

We list the phrase in the predefined template in Table 3. Except for the template introduced in previous work [2, 6], we add one more phrase, "a 3D model of {}" in the template, and while training the model, we replace "{}" with the tag of 3D shapes.

Phrases		
"a 3D model of {}.",	"a point cloud model of {}.",	"There is a {} in the scene.",
"There is the {} in the scene.",	"a photo of a {} in the scene.",	"a photo of the {} in the scene.",
"a photo of one {} in the scene.",	"itap of a {}.",	"itap of my {}.",
"itap of the {}.",	"a photo of a {}.",	"a photo of my {}.",
"a photo of the {}.",	"a photo of one {}.",	"a photo of many {}.",
"a good photo of a {}.",	"a good photo of the {}.",	"a bad photo of a {}.",
"a bad photo of the {}.",	"a photo of a nice {}.",	"a photo of the nice {}.",
"a photo of a cool {}.",	"a photo of the cool {}.",	"a photo of a weird {}.",
"a photo of the weird {}.",	"a photo of a small {}.",	"a photo of the small {}.",
"a photo of a large {}.",	"a photo of the large {}.",	"a photo of a clean {}.",
"a photo of the clean {}.",	"a photo of a dirty {}.",	"a photo of the dirty {}.",
"a bright photo of a {}.",	"a bright photo of the {}.",	"a dark photo of a {}.",
"a dark photo of the {}.",	"a photo of a hard to see {}.",	"a photo of the hard to see {}.",
"a low resolution photo of a {}.",	"a low resolution photo of the {}.",	"a cropped photo of a {}.",
"a cropped photo of the {}.",	"a close-up photo of a {}.",	"a close-up photo of the {}.",
"a jpeg corrupted photo of a {}.",	"a jpeg corrupted photo of the {}.",	"a blurry photo of a {}.",
"a blurry photo of the {}.",	"a pixelated photo of a {}.",	"a pixelated photo of the {}.",
"a black and white photo of the {}.",	"a black and white photo of a {}",	"a plastic {}.",
"the plastic {}.",	"a toy {}.",	"the toy {}.",
"a plushie {}.",	"the plushie {}.",	"a cartoon {}.",
"the cartoon {}.",	"an embroidered {}.",	"the embroidered {}.",
"a painting of the {}.",	"a painting of a {}."	

Table 3: Predefined templates for building shape-image-text pairs. Note that "{}" will be replaced by tags of the 3D shape during training.

D Visualization for image/shape retrieval

Benefiting from the alignment among 3D shapes, images, and texts via contrastive learning, our model can measure the similarity between 3D shapes and images. Therefore, our model could retrieve 3D shapes from the database given a query image. Specifically, given a query image, our model travels through the database and computes the similarity between the image and each 3D shape, where the similarity reflects the visual alignment between the image and the 3D shape. We show visual results in Figure 6, where the golden model is the 3D shape most similar to the query image.

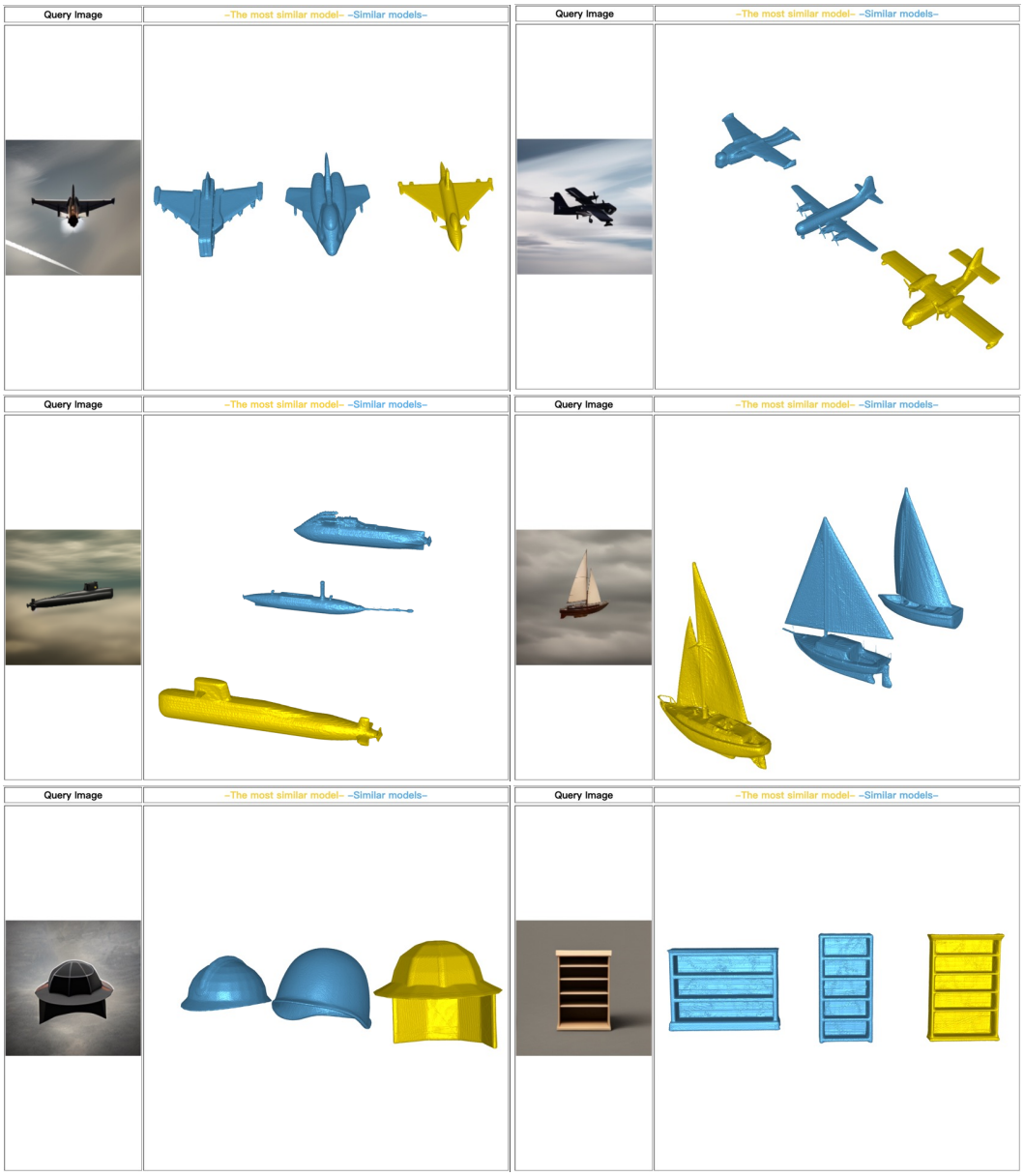

Figure 6: **3D shapes retrieval**. Given a query image, our model could retrieve similar 3D shapes from the database. Results show that the visual information is close, which proves our model could capture 3D shape information aligned with image information. (Please refer to the $'supp\_retrieve/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

E More visual comparison

Image-conditioned generation. We illustrate more image-conditioned 3D shape generation examples in Figure 7. Furthermore, the result proves that our model could capture details in the image and further generate 3D shapes faithfully. Since images only propose single-view information of 3D models, our model could also imagine plausible solutions for generating complete 3D shapes.

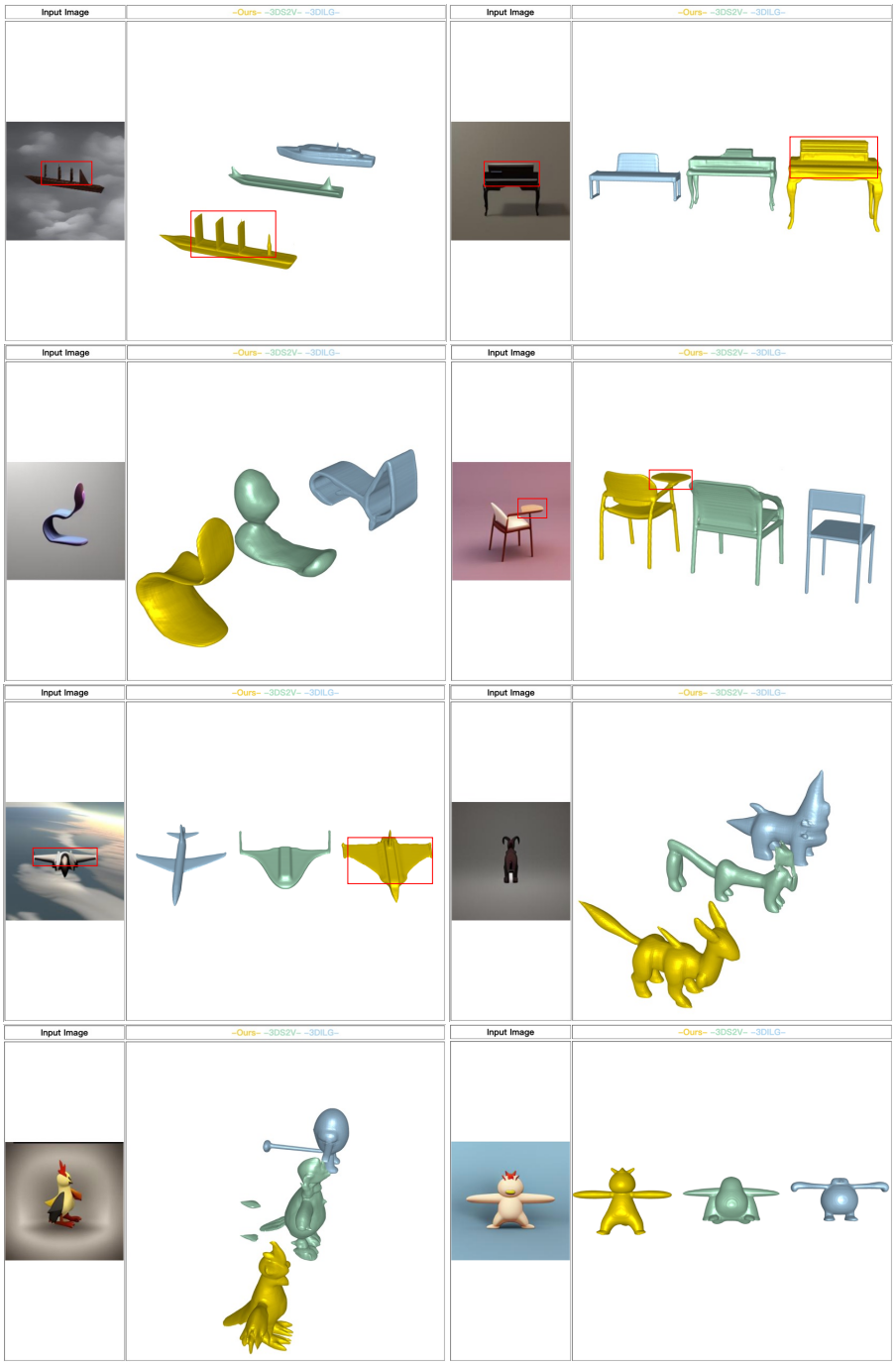

Figure 7: **Image-conditioned generation comparison:** Ours, 3DS2V [8], and 3DILG [7]. (Please refer to the $'supp\_image\_cond/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

Text-conditioned generation. We show more text-conditioned 3D shape generation results in Figure 8. According to the result, our model could understand the language correctly and map the keyword to corresponding parts in 3D shapes. The result further shows that training the model on the shape-image-text aligned space boosts the model's generative ability.

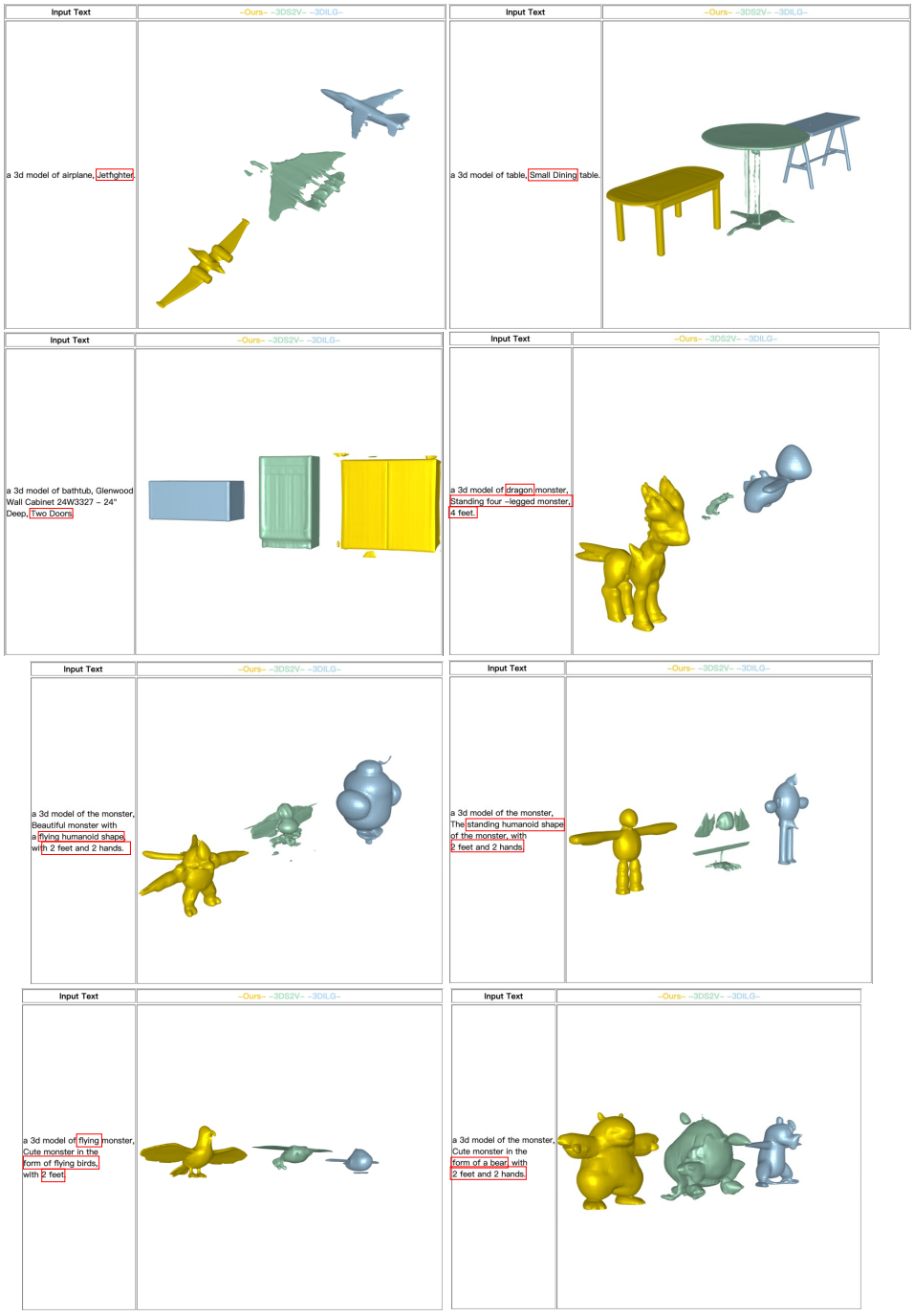

Figure 8: **Text-conditioned generation comparison:** Ours, 3DS2V [8], and 3DILG [7]. (Please refer to the $'supp\_text\_cond/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

F Test in the wild

We also test the model with data in the wild, including images from the internet and manually design text.

Conditional 3D shape generation on images from the Internet. We select some images from the Internet as conditions for the model. Results are shown in Figure 9. According to the generated 3D shapes, the model could map the visual information to 3D shapes, proving that our model could robustly handle some out-of-domain images.

Figure 9: **Conditional 3D shape generation on images from the Internet**. (Please refer to the $'supp\_wild/image/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

Conditional 3D shape generation on manual input text. Moreover, we manually design input texts as conditions for the model, and the results are shown in Figure 10. The generated 3D shapes prove that our model could capture keyword information and produce results that conform to the text.

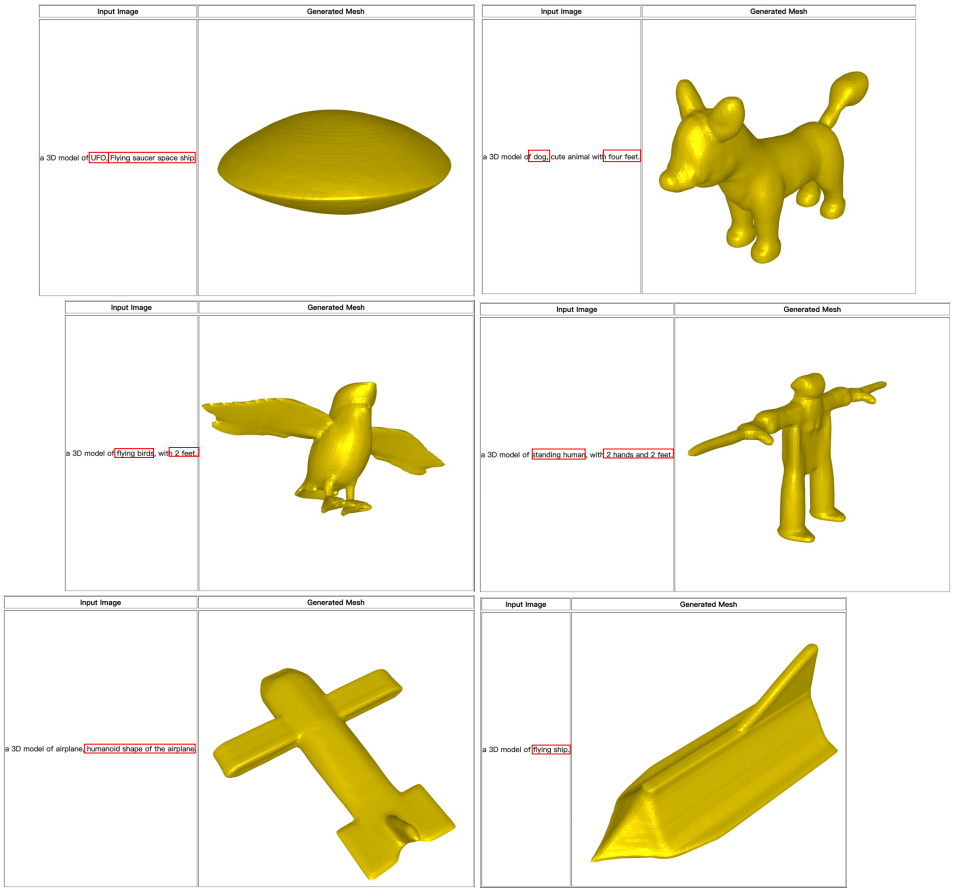

Figure 10: **Conditional 3D shape generation on manually design text**. (Please refer to the $'supp\_wild/text/*.html'$ files in the supplementary materials for the interactive 3D viewing visualization.)

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[2] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[5] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[6] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning unified representation of language, image and point cloud for 3d understanding. *arXiv preprint arXiv:2212.05171*, 2022.

[7] Biao Zhang, Matthias Nießner, and Peter Wonka. 3dilg: Irregular latent grids for 3d generative modeling. In *NeurIPS*, 2022.

[8] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *arXiv preprint arXiv:2301.11445*, 2023.