DiLightNet: Fine-grained Lighting Control for Diffusion-based Image Generation Supplementary Materials

Chong Zeng*
State Key Lab of CAD and CG,
Zhejiang University
Hangzhou, China
chongzeng2000@gmail.com

Yue Dong
Microsoft Research Asia
Beijing, China
yuedong@microsoft.com

Pieter Peers
College of William & Mary
Williamsburg, USA
ppeers@siggraph.org

Youkang Kong*
Tsinghua University
Beijing, China
kykdqs@gmail.com

Hongzhi Wu
State Key Lab of CAD and CG,
Zhejiang University
Hangzhou, China
hwu@acm.org

Xin Tong
Microsoft Research Asia
Beijing, China
xtong@microsoft.com

ACM Reference Format:
Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024. DiLightNet: Fine-grained Lighting Control for Diffusion-based Image Generation Supplementary Materials. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657396

This supplemental document contains additional DiLightNet results and information, including architecture details, a comparison to concurrent work (Lasagna [Bashkirova et al. 2023]), supplementary ablation study results, and additional results.

1 ARCHITECTURE DETAILS

DiLightNet consists of a a diffusion model that incorporates the radiance hint images using a ControlNet. Furthermore, DiLightNet also takes the original text prompt (used to generate the provisional image), and an appearance-seed as input. We desire the retain the texture and shape information from the provisional image. However, naïvely adding the provisional image as an input to the ControlNet does not work well because it is entangled with the unknown lighting from the first stage. We disentangle the relevant texture and shape information by first encoding the provisional image (with the alpha channel set to the segmentation mask). This disentanglement encoder follows Gao *et al.*'s [2020] deferred neural relighting architecture, but with a reduced number of channels to limit memory usage. In addition, we include a channel-wise multiplication between the 12-channel encoded feature map of the provisional image and the 4×3-channel radiance hints, which is subsequently passed to ControlNet. The exact details of the encoder architecture are summarized in Figure 1.

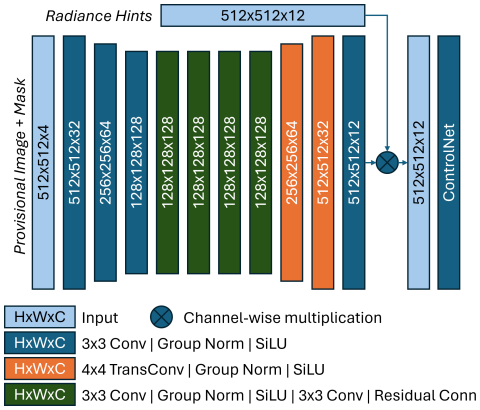


Figure 1: Provisional image encoder architecture. The output of the encoder is channel-wise multiplied with the radiance hints before passing the resulting 12-channel feature map to a ControlNet.

2 COMPARISON TO CONCURRENT WORK

Concurrent to our work, Bashkirova *et al.* [2023] introduced a lighting control method for image generation named "Lasagna". Although Lasagna shares a similar goal as DiLightNet, it uses language tokens instead of radiance hints to control the lighting and thus lacks the fine-grained lighting control of DiLightNet. Furthermore, Lasagna only supports a predefined set of 12 directional lights. Due to ambiguities in the lighting specification used in the publicly available pretrained Lasagna model, we can only compare both methods for a synthetic dataset under Lasagna's ID-0 (top) and ID-6 (front) lighting. Specifically, we perform lighting control on our synthetic test dataset, with the lighting either set as a point light source at the top or in front of the object. We then follow the same

*Work partially done during internship at Microsoft Research Asia.

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07…$15.00
https://doi.org/10.1145/3641519.3657396

Table 1: Qualitative comparison to Lasagna [Bashkirova et al. 2023].

	PSNR	SSIM	LPIPS
Ours	21.09	0.8443	0.1152
Lasagna	17.41	0.8352	0.1359

configuration as our ablation study to measure the quantitative errors using PSNR, SSIM and LIPIPS [Zhang et al. 2018]. As shown in Table 1 our method consistently outperforms Lasagna across all metrics. A qualitative comparison is shown in Figure 3.

3 ADDITIONAL ABLATION STUDY

Mask Ablation: Figure 4 shows the visual impact of passing the mask to DiLightNet. We observe that without a mask, there are more occurrences of incorrect specular highlights as the network is unable to differentiate between dark foreground pixels and background.

Number of Radiance Hints: Figure 5 shows the visual effect of using a different number of radiance hints. Using 3 radiance hints often results in missed or blurred highlights. Using too many radiance hints also tends to adversely affect the results due to the limited accuracy of the (smoothed) depth-estimated normals used for rendering the radiance hints causing sharp specular highlights to be incorrectly placed.

4 ADDITIONAL RESULTS

Examples of the synthetic test set. Figure 2 shows representative examples from the test set. Our test dataset covers a wide range of shapes with different complexities of shapes and materials.

Example of Radiance Hints: Figure 6 shows the radiance hints used by DiLightNet to control the incident lighting for a *"leather glove"*.

Additional Results: Figure 7, 8, 9, 10, 11, 12, and 13 show additional text to image generation results, including the impact of changing the content-seed using the same text prompt. For all examples, we show the results for 3 different lighting conditions.

Synthetic Results: Figure 14 shows additional results with synthetic data. The first column shows the provisional image as a reference, and the second column shows the reference image rendered under the target lighting. The last column shows the result generated under the target lighting (we select the best (lowest LPIPS) result from 4 candidate seeds). Note that our method produces plausible results that qualitatively match the reference with some minor differences in the shadows and specular highlights. These differences are mostly due to the approximate shape of the estimated depth.

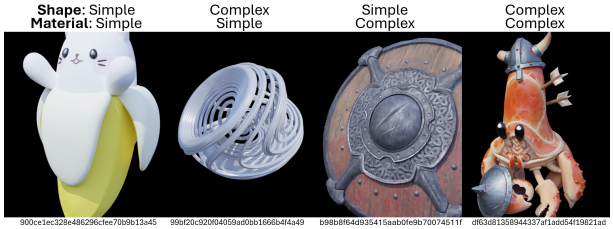


Figure 2: Representative examples, with Objaverse ID for completeness, from the synthetic test with different complexities in shape and/or material.

REFERENCES

Dina Bashkirova, Arijit Ray, Rupayan Mallick, Sarah Adel Bargal, Jianming Zhang, Ranjay Krishna, and Kate Saenko. 2023. Lasagna: Layered Score Distillation for Disentangled Object Relighting.

Duan Gao, Guojun Chen, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2020. Deferred neural lighting: free-viewpoint relighting from unstructured photographs. *ACM Trans. Graph.* 39, 6, Article 258 (2020).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. 586–595.

Figure 3: Visual comparison of DiLIghtNet with Lasagna [Bashkirova et al. 2023]. The DiLightNet result more closely matches the overall shading and shadow casted by the point light source than the Lasagna result which exhibits incorrect shadows and shading effects (e.g., on the barrel).

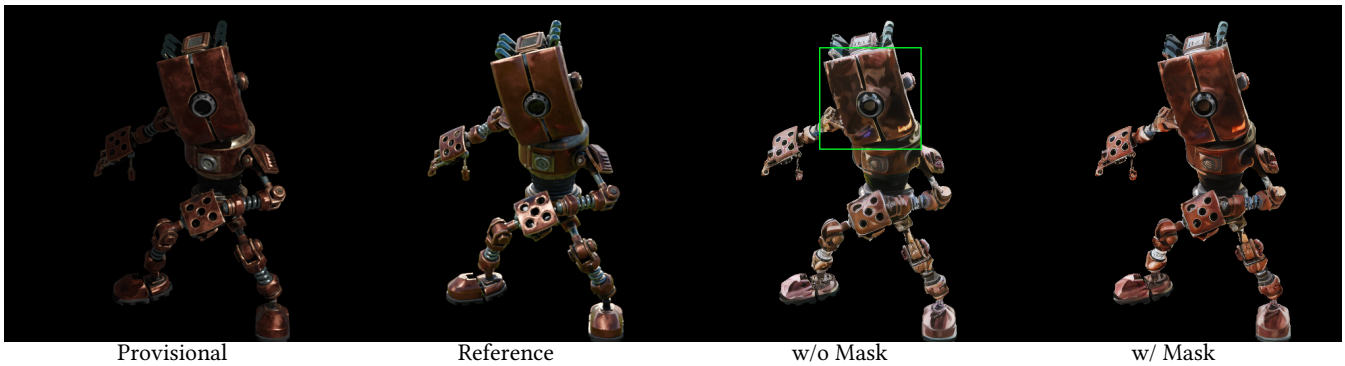


Figure 4: Not passing the mask as an extra input channel will result in more occurences of incorrect specular highlights.

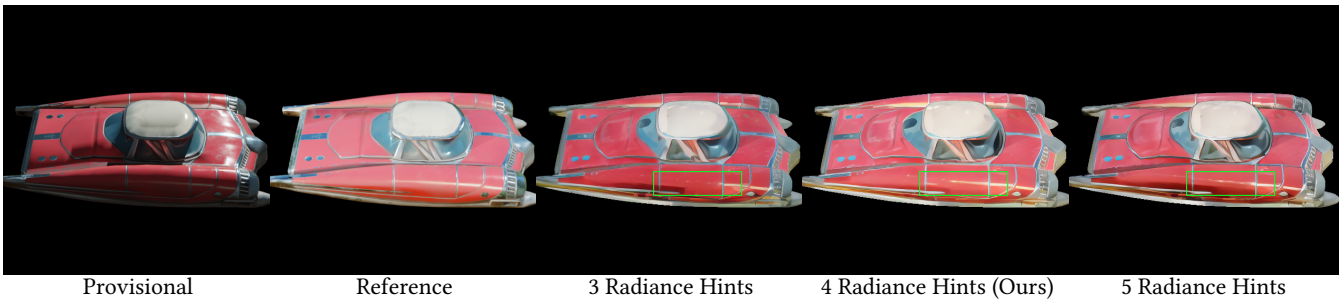


Figure 5: Ablation comparison of using a different number of radiance hints. With only *3 radiance hints*, DiLightNet misses some specular highlights, while too many hints (*5 radiance hints*) can also adversely affect results due to the inaccuracies in the depth estimates used to generate the specular radiance hints. In our implementaion we opt for using *4 radiance hints* which produces visually more plausible results.

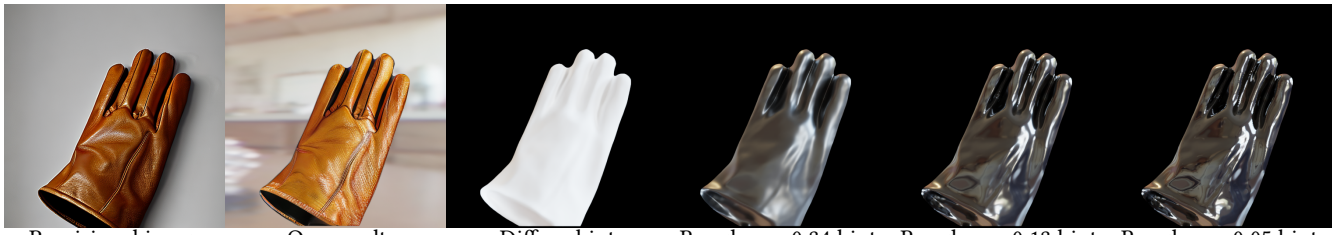


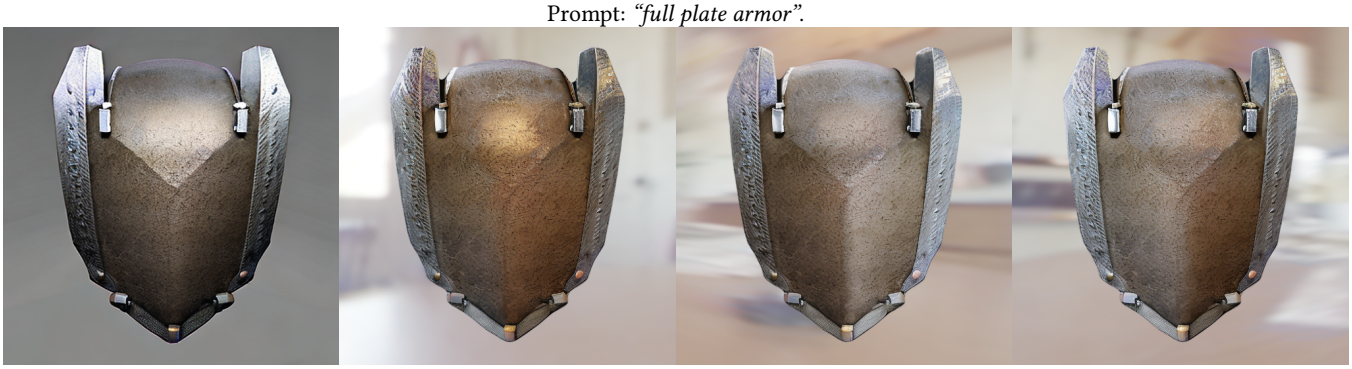
Figure 6: Example visualizations of the radiance hints for a *"leather glove"*. Note that DiLightNet leverages the learned space of images embedded in the diffusion model to generate rich shading details from the smoothed shading information encoded in the radiance hints.

Prompt: *"caterpillar work boot"*.

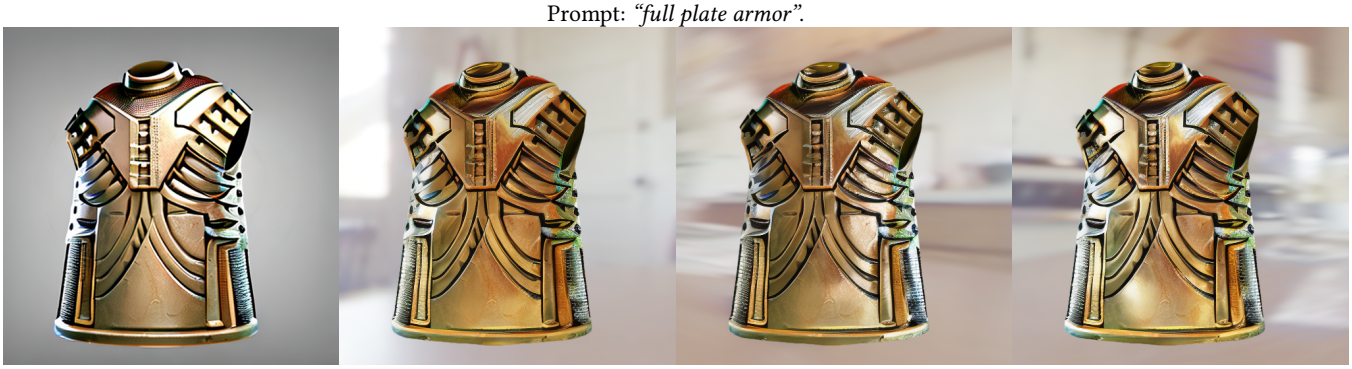
Figure 7: Text-to-image generated results with lighting control. The first column shows the provisional image as a reference, whereas the last three columns are generated under different user-specified environment lighting conditions.

Prompt: *"stone griffin"*.

Prompt: *"full plate armor"*.



Prompt: *"full plate armor"*.



Prompt: *"full plate armor"*.

Figure 8: Text-to-image generated results with lighting control. The first column shows the provisional image as a reference, whereas the last three columns are generated under different user-specified environment lighting conditions.

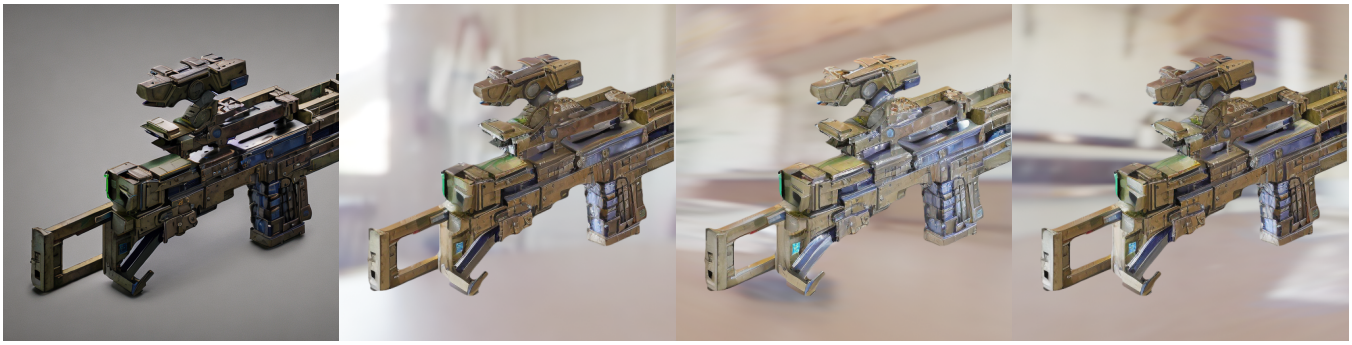
Prompt: *"leather glove"*.

Prompt: *"leather glove"*.

Prompt: *"leather glove"*.

Prompt: *"leather glove"*.

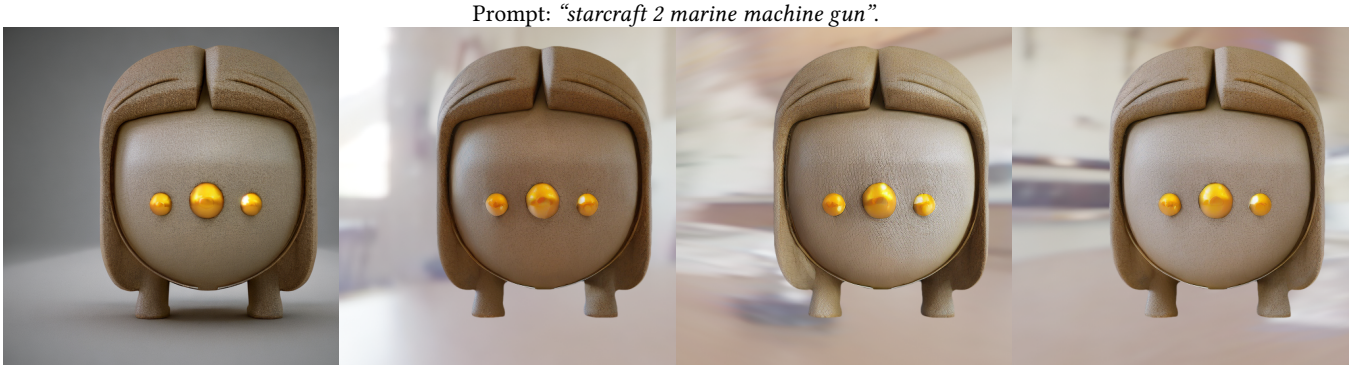
Figure 9: Text-to-image generated results with lighting control. The first column shows the provisional image as a reference, whereas the last three columns are generated under different user-specified environment lighting conditions.



Prompt: *"starcraft 2 marine machine gun"*.

Prompt: *"starcraft 2 marine machine gun"*.

Prompt: *"starcraft 2 marine machine gun"*.

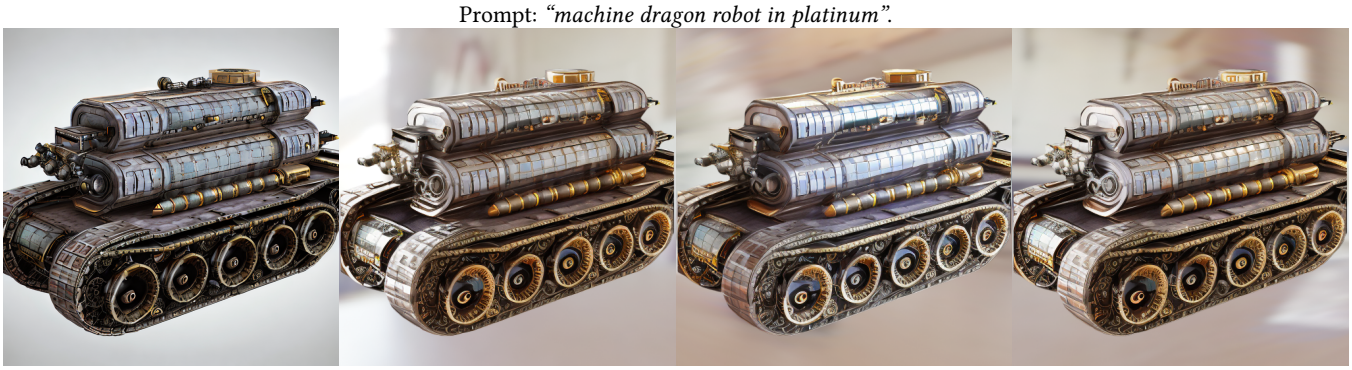


Prompt: *"3d animation character minimal art toy"*.

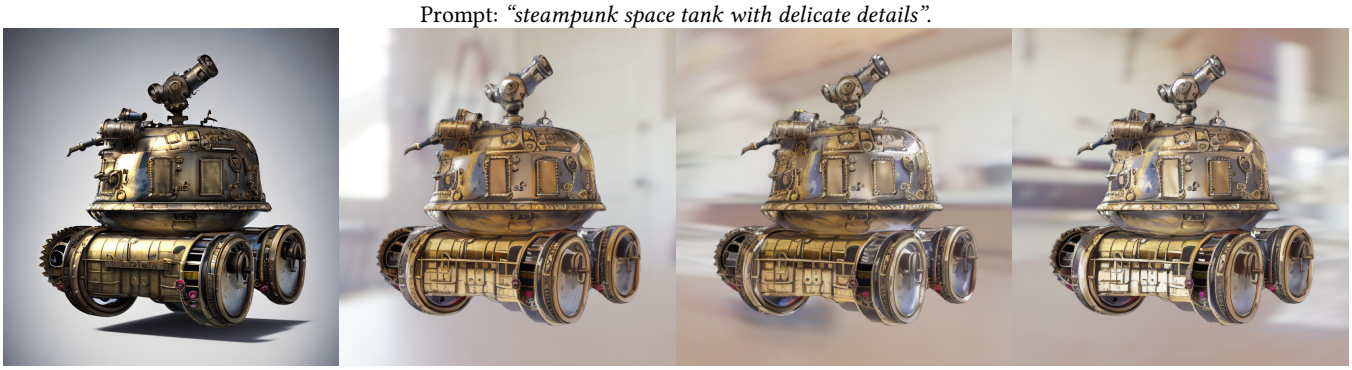
Figure 10: Text-to-image generated results with lighting control. The first column shows the provisional image as a reference, whereas the last three columns are generated under different user-specified environment lighting conditions.

Prompt: *"machine dragon robot in platinum"*.

Prompt: *"machine dragon robot in platinum"*.

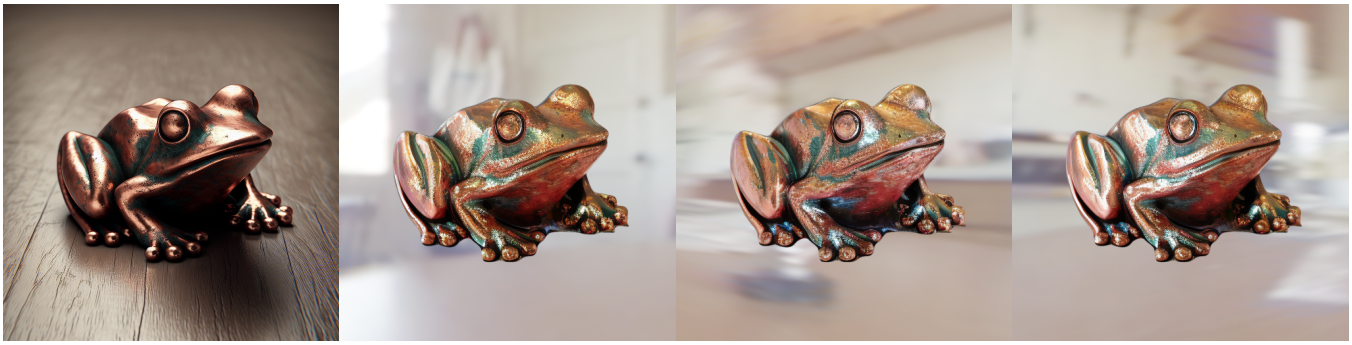


Prompt: *"steampunk space tank with delicate details"*.



Prompt: *"steampunk space tank with delicate details"*.

Figure 11: Text-to-image generated results with lighting control. The first column shows the provisional image as a reference, whereas the last three columns are generated under different user-specified environment lighting conditions.



Prompt: *"Rusty copper toy frog with spatially varying materials some parts are shinning other parts are rough".*

Prompt: *"An elephant sculpted from plaster and the elephant nose is decorated with the golden texture".*

Prompt: *"Rusty sculpture of a phoenix with its head more polished yet the wings are more rusty".*

Prompt: *"Rusty sculpture of a phoenix with its head more polished yet the wings are more rusty".*

Figure 12: Text-to-image generated results with lighting control. The first column shows the provisional image as a reference, whereas the last three columns are generated under different user-specified environment lighting conditions. The provisional images are generated with *DALL-E3* instead of *stable diffusion v2.1* to better handle the more complex prompt.

Prompt: *"A decorated plaster rabbit toy plate with blue fine silk ribbon around it".*

Prompt: *"A decorated plaster round plate with blue fine silk ribbon around it".*

Figure 13: Text-to-image generated results with lighting control. The first column shows the provisional image as a reference, whereas the last three columns are generated under different user-specified environment lighting conditions. The provisional images are generated with *DALL-E3* instead of *stable diffusion v2.1* to better handle the more complex prompt.

Figure 14: Additional results with synthetic data. The first column shows the provisional image as a reference, whereas the second column is the reference image rendered under the target lighting. The last column is the result generated by DiLightNet under the target lighting.