PhyS-EdiT: Physics-aware Semantic Image Editing with Text Description

Supplementary Material

Ziqi Cai[1,2] Shuchen Weng[3] Yifei Xia[1,2] Boxin Shi[1,2]*

[1]State Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3]Beijing Academy of Artificial Intelligence

czq@stu.pku.edu.cn, shuchenweng@pku.edu.cn, yfxia@pku.edu.cn, shiboxin@pku.edu.cn

A. Implementation Details

Denoising networks. We initialize the networks U_{low} and U_{high} with pretrained weights from InstructPix2Pix [1]. U_{low} further incorporates a ControlNet [7], with weights initialized from the baseline U-Net model. Following the protocol in [6], we employ an auxiliary encoder. The encoded input image is element-wise multiplied with physical conditions before being fed into the network to enhance generalization.
Fusion network. The fusion network employs a Convolutional Neural Network (CNN) as its backbone, operating directly in the latent space. This allows the model to learn more diverse and disentangled representations for both physical and semantic editing.
Data rendering. We render images using Blender 4.2 [2] with the Cycles renderer at a resolution of 1024×1024 and a sample count of 64. During training, these images are resized to 512×512. To ensure consistency, we normalize the scenes such that the object is centered and fully visible.

B. Baseline Configurations

InstructPix2Pix (IP2P) [1]. We employ IP2P [1] as a baseline for both material and semantic editing. We utilize the official code release and pretrained weights. For material editing, we adhere to the methodology in [4], providing the following instructions to the model:

- **Roughness:** Make the {object} more/less shiny.
- **Metallicity:** Make the {object} more/less metallic.
- **Albedo:** Make the {object} more/less gray.
- **Transparency:** Make the {object} more/less transparent.

For semantic editing, we utilize prompts consistent with the IP2P dataset [1].
Subias et al. [5]. We deploy the official code release and pretrained weights from this model, which only supports the adjustment of roughness and metallicity.
DiLightNet [6]. We utilize the official code release and pretrained weights. The model supports lighting control, but does not allow material editing, leading to variations in the editing results based on the appearance seed.

*Corresponding author.

Stable Diffusion 3 [3]. We use the medium inpaint version of Stable Diffusion 3 for semantic editing. To guide the model towards the intended editing effects, we use the editing instructions as described in IP2P [1].

C. Dataset Visualization

The PR-TIPS dataset includes pairwise images with varying levels of roughness, metallicity, albedo, and transparency under diverse lighting setups. To provide an overview of the diversity and quality of our dataset, we present examples of image-target pairs used in our experiments. Figure D illustrates the variety of materials, lighting conditions, and objects in the dataset.

D. Additional Results

D.1. Generalization

We present additional real-image results in Fig. A to show our model's resistance to overfitting.

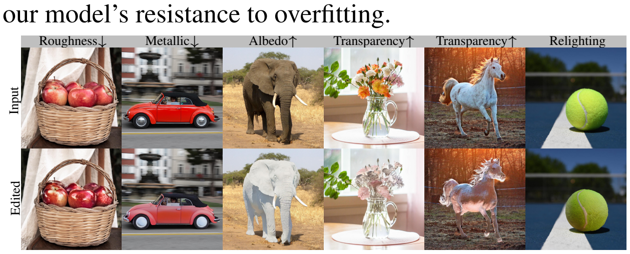

Figure A. Real-image results.

D.2. Retraining IP2P

We retrain IP2P [1] on our PR-TIPS dataset for material editing. The results are shown in Fig. B.

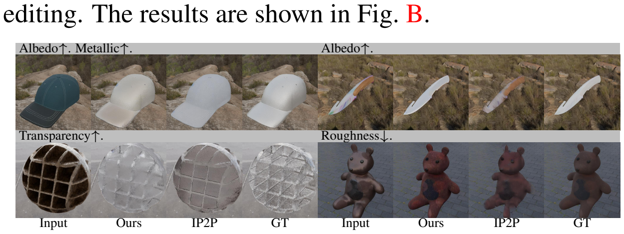

Figure B. Comparison to retrained IP2P.

D.3. Influence of Pre-trained Models

Pre-trained models are usually reliable but may struggle in challenging scenarios like translucent objects or dark scenes, causing minor deviations in physical edits. Examples of such failures are shown in Fig. C. Despite inaccuracies in low-level features, the high-level network maintains semantic robustness.

Figure C. Impact of on pretrained model results.

D.4. Additional Qualitative Results

We present the complete visualization of the Fig. 3 in Fig. E and Fig. F. Additional comparison results are presented in Fig. G and Fig. H. As observed, our method consistently generates high-quality results.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 1

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024. 1

[4] Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, Bill Freeman, and Mark Matthews. Alchemist: Parametric control of material properties with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[5] J Daniel Subias and Manuel Lagunas. In-the-wild material appearance editing using perceptual attributes. In *Computer Graphics Forum*, 2023. 1

[6] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. DilightNet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH Conference Papers*, 2024. 1

[7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, 2023. 1

Figure D. Examples from our dataset, showcasing the editing prompts, input images, and the corresponding output target.

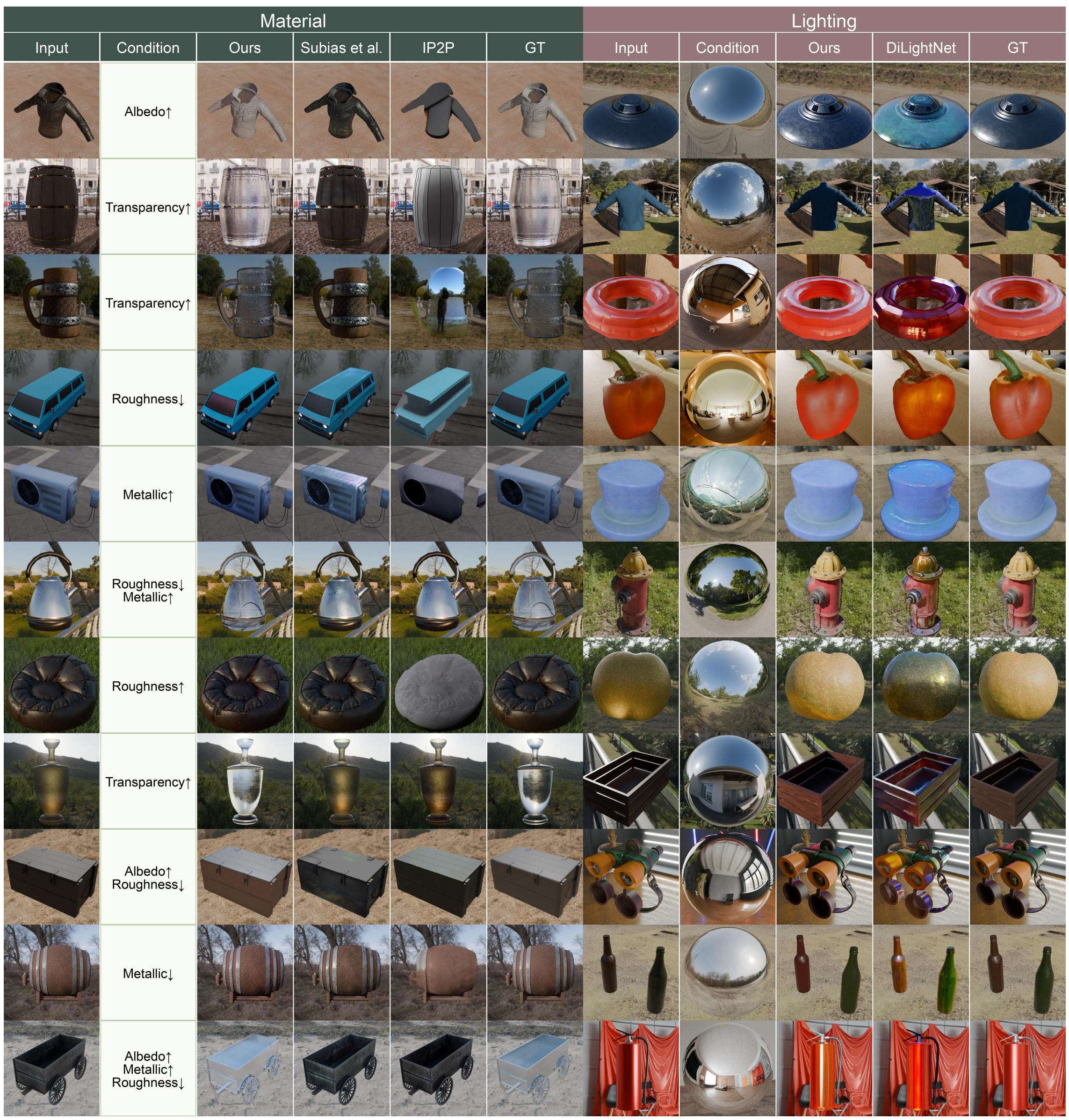

Figure E. The complete visualization for material editing and lighting editing, including input, condition, output, and ground truth.

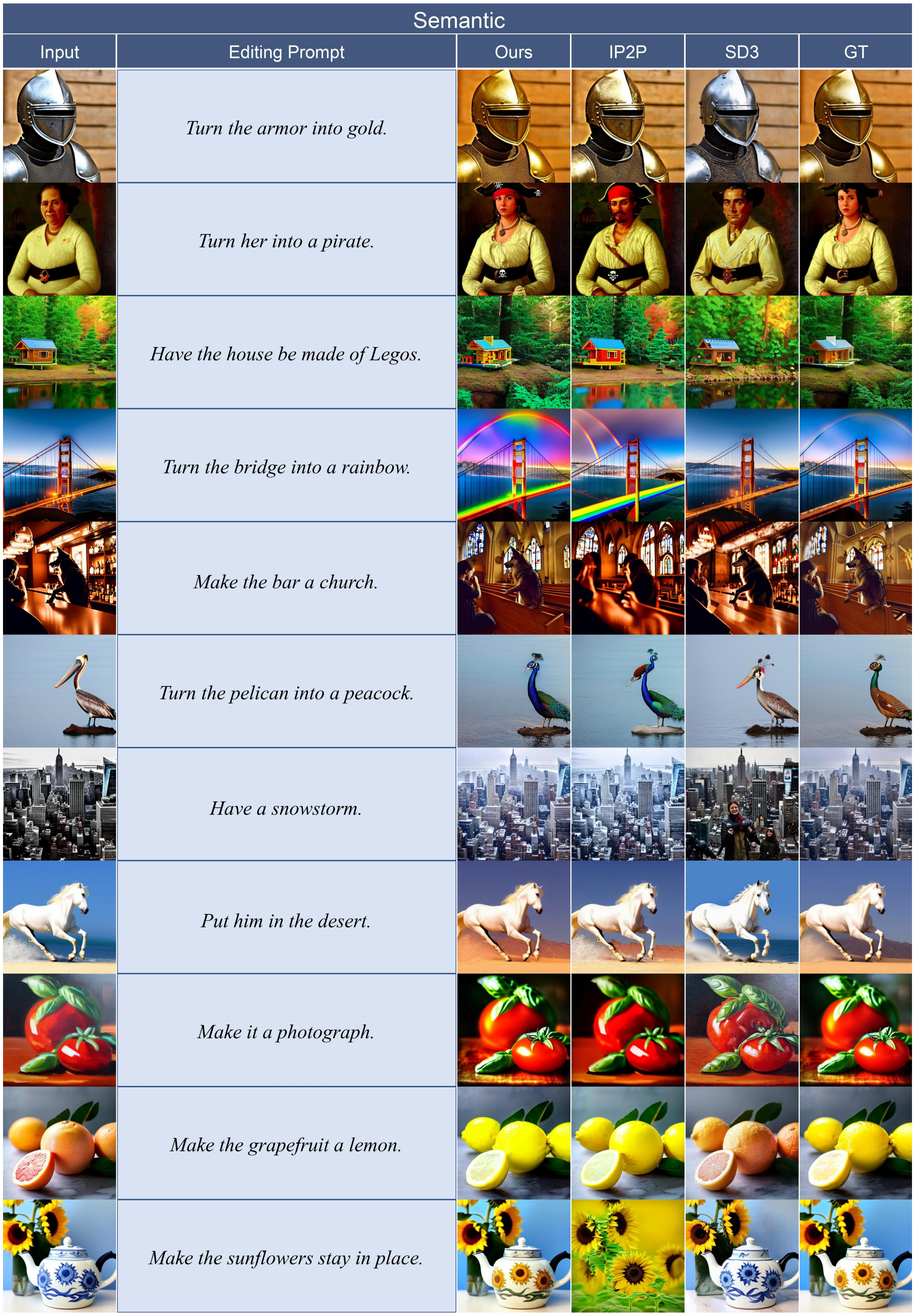

Figure F. The complete visualization for semantic editing, including input, condition, output, and ground truth.

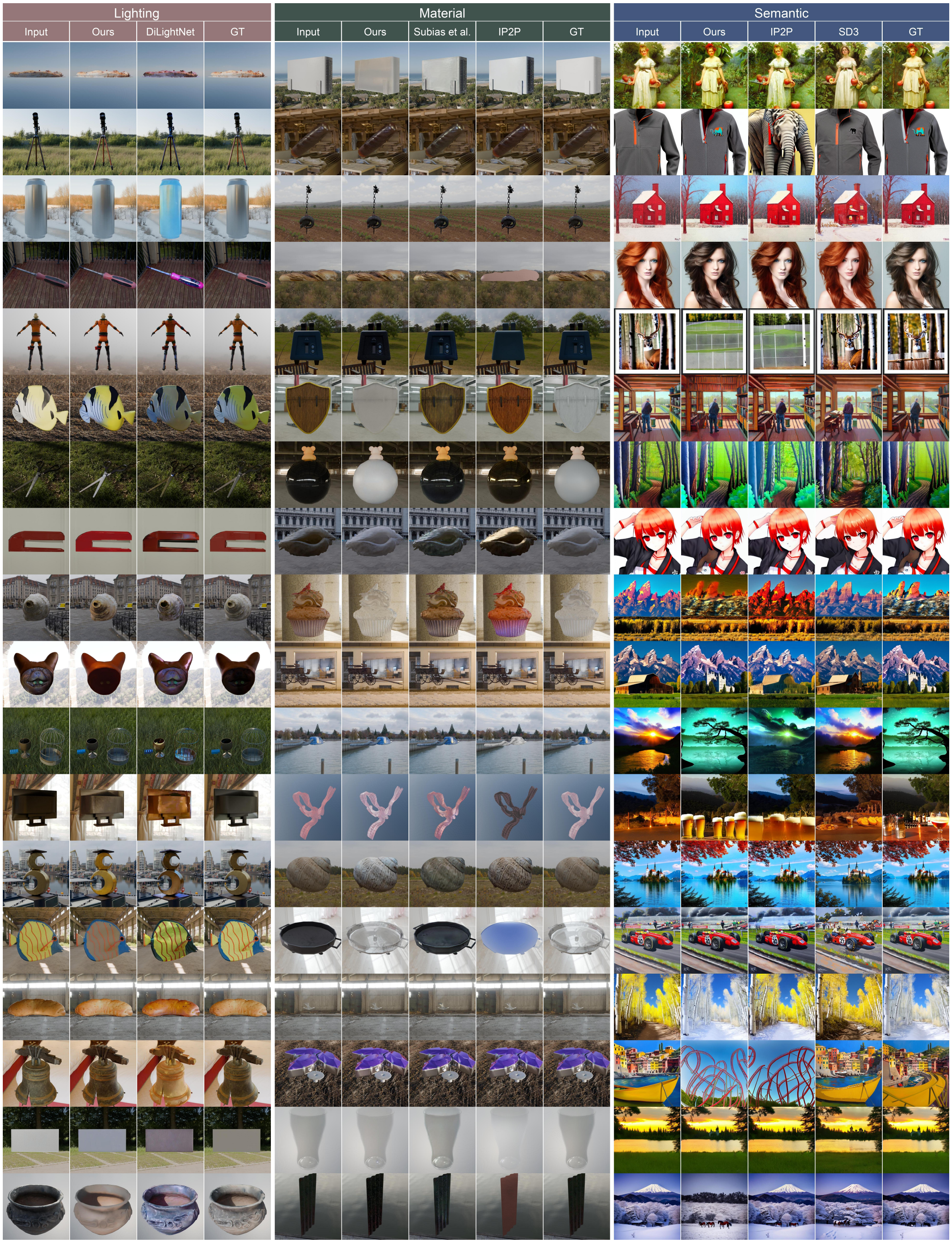

Figure G. Additional comparison results for material, lighting, and semantic editing (specific conditions omitted for clarity).

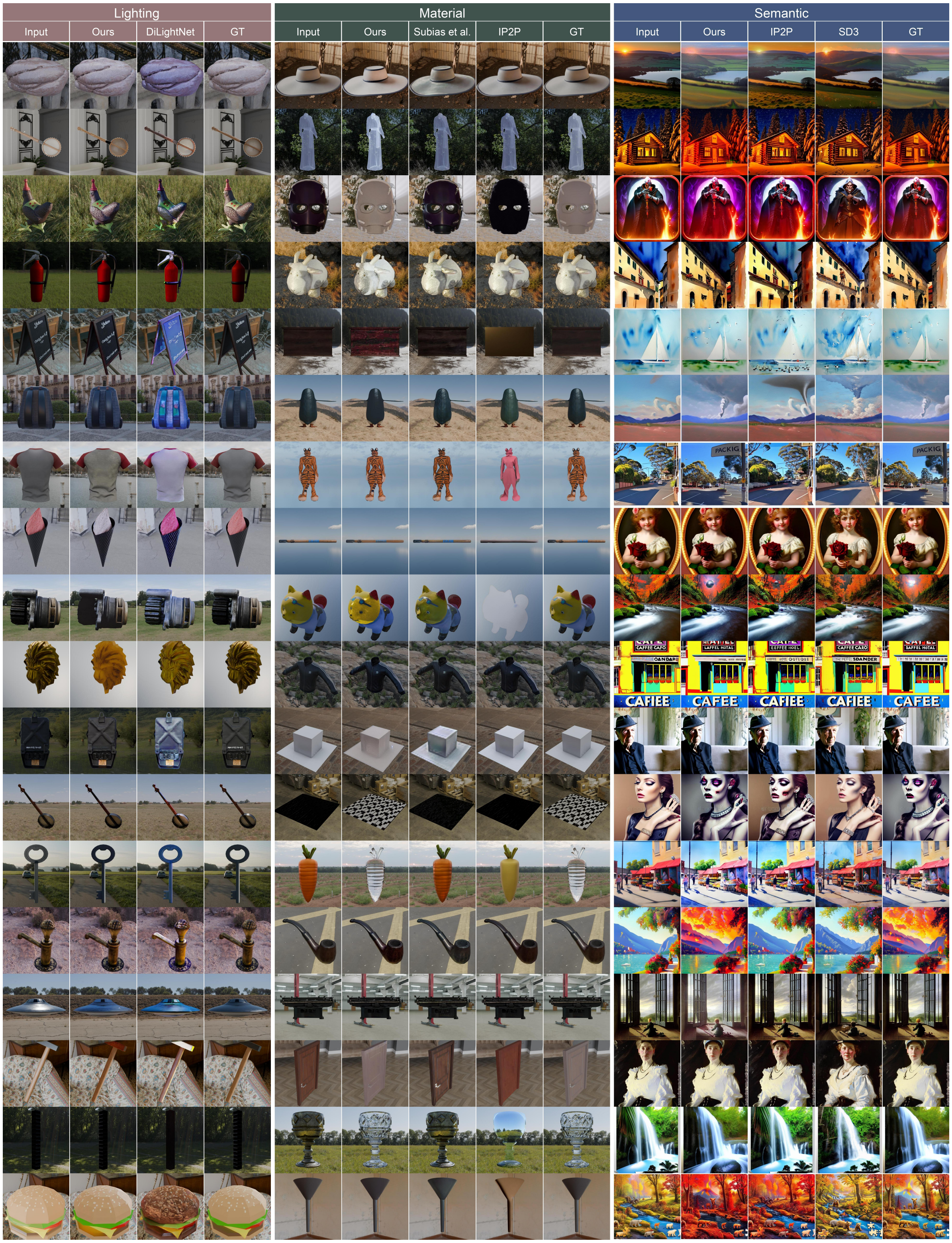

Figure H. Additional comparison results for material, lighting, and semantic editing (specific conditions omitted for clarity).