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

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

Figure 1. Illustration of PhyS-EdiT's capabilities in disentangled control over lighting, material properties, and high-level semantics. The symbols ↑ and ↓ indicate an increase and a decrease in property value, respectively. Rows show variations in lighting environments, while columns show precise adjustments to individual material properties. The rightmost column presents a semantic edit with the prompt: *"The knight is holding the sword."*

Abstract

Achieving joint control over material properties, lighting, and high-level semantics in images is essential for applications in digital media, advertising, and interactive design. Existing methods often isolate these properties, lacking a cohesive approach to manipulating them simultaneously. We introduce PhyS-EdiT, a novel diffusion-based model that enables precise control over four critical material properties: roughness, metallicity, albedo, and transparency while integrating lighting and semantic adjustments within a single framework. To facilitate this disentangled control, we present PR-TIPS, a large and diverse synthetic dataset designed to improve the disentanglement of material and lighting effects. PhyS-EdiT incorporates a dual-network architecture and robust training strategies to balance low-level physical realism with high-level semantic coherence, supporting localized and continuous property adjustments. Extensive experiments demonstrate the superiority of PhyS-EdiT in editing both synthetic and real-world images, achieving state-of-the-art performance on material, lighting, and semantic editing tasks.

*Corresponding author.

1. Introduction

Text-based image editing models [3, 10, 28, 34, 51] have demonstrated significant success in various professional visual content creation applications, including business design, advertising, and entertainment. Given a text description to provide the editing instruction, these models can effectively manipulate high-level image semantics (*e.g.*, category, pose, and layout), reducing the time and technical expertise required by users.

Complementing the high-level "what" and "where" of the scene, the low-level physical properties (*e.g.*, material, and lighting) are essential to realistically render "how" the visual appearance looks like. Recent research explores physical-aware image editing using pairwise rendered training data and incorporating physical principles. These studies have achieved separate control over material [38, 40] or lighting [4, 47, 48, 52].

However, despite these advances, two primary challenges remain: *(i)* **Incomplete material-lighting disentanglement.** The complex interplay between material and lighting often results in incomplete editing when only one aspect is considered [38, 52], limiting the control capability needed for precise adjustments and hindering the accurate rendering of low-level physical properties. *(ii)* **Limited se-**

mantic editing. Current physics-aware editing approaches focus primarily on material and lighting control, overlooking the alignment with high-level image semantics [38, 47]. It is still challenging to balance between physical realism and semantic consistency.

In this paper, we propose **PhyS-EdiT**, a unified framework for **Phy**sics-aware **S**emantic image **Edi**ting with **T**ext descriptions that leverages diffusion models for precise manipulation of material properties, lighting, and high-level semantics. As demonstrated in Fig. 1, our versatile approach allows for modifications ranging from varying lighting conditions (first column) to adjustments of metallic levels, roughness, transparency, albedo, and semantics (subsequent columns), producing consistent and realistic results. Specifically, for low-level physical properties, our PhyS-EdiT supports fine-grained disentangled control over lighting and material properties. These properties are represented as a joint feature maps, which guide a conditional network in injecting such low-level physical knowledge into the diffusion models. For high-level semantics, our PhyS-EdiT combines text-based instructions with accurate physical conditions using a fusion network, maintaining the full capability of semantic editing.

For effective training and evaluating relevant editing models, we create and collect the **PR-TIPS**, a large-scale dataset of **P**hysical **R**endering-based **T**ext and **I**mage **P**air **S**et. PR-TIPS contains an extensive collection of 3D-rendered scenes with systematically varied lighting and material properties, providing a rich source of paired data for disentangled control over physical properties. Our contributions can be summarized as follows:

- We propose a unified framework for controllable editing of both disentangled low-level physical properties and flexible high-level image semantics.
- We develop a robust training strategy that minimizes the ambiguity between low-level and high-level editing goals, achieving versatility without compromising precision.
- We present PR-TIPS, a large-scale dataset tailored for disentangled control over physical properties, providing a rich source of paired data for training and evaluation.

2. Related work

2.1. Diffusion Models for Image Generation

Diffusion models [18, 39] iteratively denoise a Gaussian distribution to generate images, learning to reverse the forward process where Gaussian noise is progressively added to an image. Recent advances have led to their dominance in image generation, outperforming other generative models [16, 22]. To further reduce computing cost and generate high-resolution images, Stable Diffusion [34] is designed. Its success inspires subsequent image generation works [9, 31], and image editing methods discussed later.

2.2. Text-driven Semantic Image Editing

Text-driven image editing has gained significant research interest, as text descriptions offer an intuitive and flexible means to guide image manipulations. Early work [46, 49] focuses on controlling image category via text guidance. Subsequent research [26, 50] explores effective methods for injecting text features into image generation models. The emergence of pre-trained text-image models like CLIP [32] enables the open-vocabulary editing approaches with promising applications [6, 25]. The large-scale image-text datasets [36, 37] further motivate the development of high-quality, open-source text-to-image models [35]. Leveraging the generative priors of these models, recent advancements have been made in super resolution [15, 41], image colorization [44, 45], and reflection removal [19, 54]. Despite substantial progress in high-level semantic editing, text-driven methods still struggle to comprehend and accurately manipulate low-level physical properties, such as lighting and material. This limitation often results in physically unrealistic outputs when precise control over such properties is needed.

2.3. Physics-aware Image Property Editing

Physics-aware image property editing, which includes material editing and lighting editing, confronts the highly ill-posed challenge of decomposing these elements from a single image. Early explorations [8, 12, 13, 29] utilize empirical formulas to approximate the separation of material and lighting properties. Recent advancements in diffusion models have demonstrated their potential in effectively separating material and lighting components. Notably, the recent works [1, 4, 27, 48] have demonstrated the powerful capabilities of diffusion models, producing impressive results in editing physical properties. However, these approaches typically focus on either lighting [47, 52] or material [38, 40], struggling to provide a comprehensive understanding of the full spectrum of physical properties inherent in images. The gap between high-level semantic editing and low-level physics-aware editing highlights the pressing need for an approach that effectively integrates these properties into a joint framework.

3. Dataset

Although previous work [38] collects a dataset with basic material variations (*e.g.*, roughness, metallicity, albedo, and transparency), it lacks pixel-aligned maps for fine-grained local editing and pairwise data for disentangled control of material and lighting. Similarly, DiLightNet [47] provides a large dataset with diverse lighting conditions, but it does not support material editing. To address these limitations, we introduce PR-TIPS, a dataset of **P**hysical **R**endering-based **T**ext and **I**mage **P**air **S**et, designed to scale up the diversity

and scope of scenes and objects. PR-TIPS provides extensive pairwise data, including pixel-aligned images and text-image pairs with varied instructions, supporting more fine-grained disentangled control over both material and lighting adjustments.

Material diversity. We collect 3D assets from a subset of Objaverse [7], selecting 13K models that possess Principled BSDF material properties. To further adjust material properties continuously, we integrate three ShaderNodeMath nodes, applying random values to properties (*e.g.*, Metallic, Roughness, and Transmission Weight). Additionally, we use a ShaderNodeMixRGB node to blend albedo color variations, supporting fine-grained color transitions. Rendering is performed in Blender's Cycles renderer [5], ensuring precise physical interaction with lighting.

Lighting diversity. To achieve lighting diversity, we utilize over 500 HDRI environment maps from Polyhaven[1], applying random rotations to each to simulate diverse lighting conditions. Camera configurations are randomized for field of view, distance, and position around the scene origin, allowing a wide range of perspectives. This setup generates a final dataset, including 300K rendered images to provide substantial variation in materials, lighting, and viewpoints to support robust training.

Text annotation process. We employ a refined text annotation strategy using GPT-4o [11], a Multi-modal Large Language Model (MLLM) with Chain of Thought (CoT) prompting techniques [43]. The model generates detailed text descriptions for both the original and edited images in three structured steps: *(i)* Identify the scene context (*e.g.*, day or night, indoor or outdoor) and the object type, *(ii)* Infer the object's material properties, leveraging knowledge of common materials and typical lighting effects, and *(iii)* Formulate an editing instruction based on the object type, material attributes, and lighting adjustments.

Dataset statistics. PR-TIPS consists of 300K pairs, with 298K for training and 2K for evaluation. To improve data diversity and promote model generalization, we apply various random augmentations to each sample on the fly during training. Specifically, we use random rotations within the range $r \in [-30°, 30°]$, scaling factors $s \in [0.9, 1.1]$, and shear transformations $sh \in [-10°, 10°]$. The translations are applied in both horizontal and vertical directions, with the offsets t_x and t_y drawn from the range $[-0.1, 0.1] \times$ image dimensions. Additionally, horizontal and vertical flips are applied with a probability of $p = 0.5$. These augmentations are intended to improve the robustness and generalization of the model by simulating a wide range of transformations during training.

Summary. PR-TIPS thus provides an extensive and versatile dataset for physics-aware image editing, supporting disentangled control over material and lighting properties.

[1] https://polyhaven.com

4. Methodology

This section begins with an overview of our framework in Sec. 4.1. Next, we discuss how our model extracts and injects physical properties in Sec. 4.2. Finally, we integrate physical knowledge to image semantic editing models in Sec. 4.3, supporting both low-level detail fidelity and high-level semantic consistency.

4.1. Overview

As illustrated in Fig. 2, we propose PhyS-EdiT, a physics-based framework to improve text-based image editing with physical awareness.

Physical embedding. The physical properties, *i.e.*, lighting conditions and material conditions, are represented as condition maps C_L and C_M, respectively. These structured condition maps offer spatially aligned hints about the physical world, providing our model with robust priors for lighting and material properties.

Forward process. In the forward process of our diffusion model [18, 39], the latent variable z_t at each timestep t is represented as a linear combination of the initial latent z_0 and Gaussian noise ϵ:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t}\epsilon. \tag{1}$$

Denoising networks. Our model incorporates a dual-denoising network design, composed of a low-level editing network U_{low} and a high-level editing network U_{high}. The low-level editing network U_{low}, structured as a U-Net [3], integrates a new physical condition injection module to guide fine-grained adjustments based on material and lighting conditions. Concurrently, the high-level editing network U_{high} performs broader semantic modifications. Finally, a fusion network U_{fusion} combines the outputs of U_{low} and U_{high}, producing a refined result that balances physical condition editing with global semantic consistency. This dual-network approach enables the editing of precise physical conditions without compromising the ability of the model to make coherent semantic editings.

Backward process. The backward process involves iterative denoising, where the dual networks U_{low} and U_{high} gradually refine the latent code. The denoising networks U_{low} and U_{high} first independently process the noisy latent code $\hat{z}_t$ according to their respective roles. At each timestep t, the fusion network U_{fusion} then combines the outputs from U_{low} and U_{high} to yield the next-step latent code:

$$\hat{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{z}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} U_{\text{fusion}}(\hat{z}_t, t) \right) + \sigma_t \epsilon, \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ denotes the accumulated noise factor, and σ_t is the standard deviation of the noise added at each step. Through T iterations, the model refines the latent code to produce $\hat{z}_0$, achieving a balance between low-level physical editing and high-level semantic coherence.

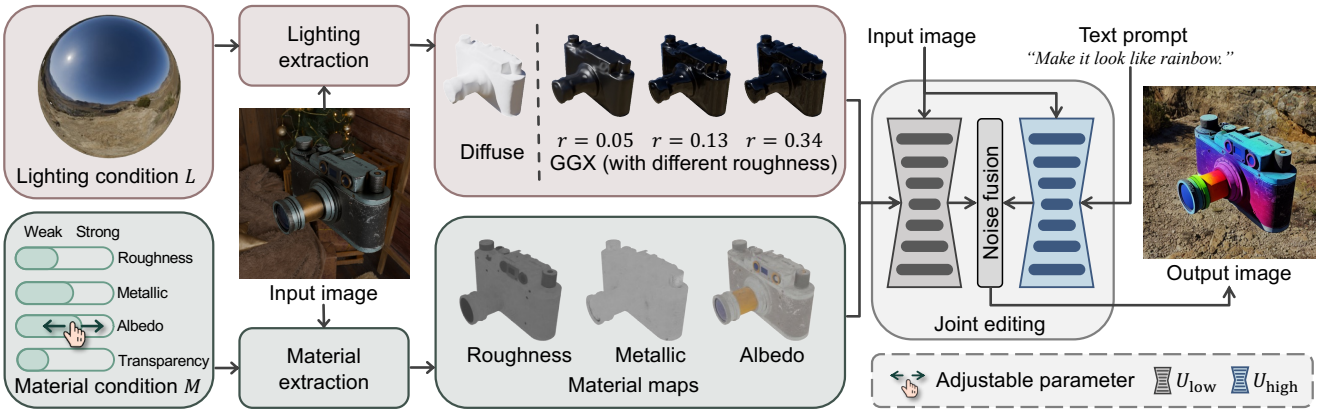

Figure 2. Pipeline of PhyS-EdiT. Given an input image with the desired lighting condition L and material condition M, the model first uses the input image and L to generate lighting condition maps and, concurrently, the input image and M to generate material maps. These maps condition the low-level editing network U_{low}. Simultaneously, a semantic text prompt modulates the high-level editing network U_{high}. The outputs from U_{low} and U_{high} are then combined through the noise fusion module, resulting in the final edited image.

4.2. Material-lighting Joint Representation

While previous generative models [3, 34] excel at generating visually appealing images, they often fall short in adhering to real-world physical principles, such as accurate lighting dynamics and material interactions (*e.g.*, realistic reflections). This limitation arises from their inability to interpret or integrate physical properties into the generative process. In this section, we introduce a novel approach that incorporates physical properties into image editing models, enabling physics-aware image editing.

Physical property extraction. We categorize physical properties into two essential components: material and lighting. Previous approaches often focus on controlling either material or lighting independently; however, our goal is to achieve unified control over both. The primary challenge involves representing these properties in a format that diffusion models can interpret effectively. Traditional input formats, such as text prompts for material properties or environment maps for lighting conditions, often fail to maintain spatial alignment, thereby complicating the model's ability to understand and utilize the physical context effectively.

To address this, we leverage pretrained models [48] to extract well-aligned physical property maps from the input image I. We define the material condition map C_M as $C_M = [A, R, M, T]$, where A, R, M, and T represent albedo, roughness, metallicity, and transparency maps, respectively. For the lighting condition map C_L, we utilize an off-the-shelf depth estimator [2] to reconstruct a coarse geometry of the scene following [47]. This geometry is then rendered under predefined proxy materials using environment maps. We define the lighting condition as follows:

$$C_L = f_{\text{lighting}}(I, E), \tag{3}$$

where f_{lighting} represents the function that determines the lighting maps based on the input image I and environment map E. Finally we combine the lighting condition map C_L with the material condition map C_M to form the physical condition map C_{phys}:

$$C_{\text{phys}} = C_M \oplus C_L, \tag{4}$$

where $\oplus$ denotes the channel-wise concatenation. By providing the model with material-lighting joint inputs C_{phys}, we bridge the representation gap between different physical properties.

Physical embedding injection. With available structured representations of the material and lighting conditions, the next step is to inject these conditions into the diffusion model. We achieve this by encoding the input image I using an encoder $E(\cdot)$ inspired by the deferred neural relighting architecture [14]. The resulting encoded feature map $F = E(I)$ captures richer information. To capture interactions between material and lighting, we perform a channel-wise multiplication between the encoded features and the combined physical conditions:

$$C_{\text{emb}} = F \odot C_{\text{phys}}, \tag{5}$$

where $\odot$ denotes element-wise multiplication. This operation allows the model to learn complex relationships between the physical properties and the visual appearance of the image. The resulting physical embedding C_{emb} is then used as a conditioning input for ControlNet [51], which integrates this information into the initial feature spaces of the diffusion model.

$$\hat{z}_{t-1} = \text{ControlNet}(U_{\text{low}}, \hat{z}_t, t, C_{\text{emb}}). \tag{6}$$

This injection process effectively captures both local details and high-level contextual features, culminating in a compact, physics-aware embedding that meets the conditioning requirements of the diffusion model.

4.3. Bridging Physics to Semantic Image Editing

Effective image editing models should seamlessly integrate both high-level semantic understanding and detailed physical property manipulations. Previous training strategies, such as fine-tuning attention layers [24] and using low-rank adaptations like LoRA [21], often struggle to preserve the original capabilities of the pre-trained model while adapting to new tasks. These methods can inadvertently weaken the model's ability to maintain generative priors, leading to suboptimal performance in scenarios that require fine-grained control over image properties, such as material textures and lighting effects. To address these challenges, we introduce the following strategies:

Preserving generative priors for semantic consistency. To maintain the semantic coherence and editing quality of the pre-trained diffusion model, we initially freeze the weights of the U-Net architecture, $\theta_{\text{U-Net}}$, for the first 30K iterations. This strategy preserves the foundational generative capabilities of the model during early training. Following this phase, we selectively unfreeze the U-Net's weights, keeping the decoder frozen, and continue training for an additional 15K iterations. This controlled fine-tuning improves the model's capacity to incorporate new conditioning inputs effectively while ensuring high-quality semantic consistency and robust generative performance.

Multitask supervision for learning diverse edits. We employ a multitask learning framework that leverages paired datasets for robust training. Our model is trained on an equal mix of our PR-TIPS synthetic dataset and the high-level editing dataset from [3], with a 1:1 ratio. To improve stability, we introduce a training augmentation where, at times, the edited image is identical to the input image, helping the model learn to preserve stability when no edits are required.

Sequential training. Our training initiates with material editing, focusing on mastering material representations under consistent lighting conditions. We quantify the material editing loss as:

$$\mathcal{L}_{\text{material}} = \mathbb{E}_{I,C_M}\left[||\epsilon - \epsilon_\theta\left(z_t, t, C_M\right)||_2^2\right], \tag{7}$$

where C_M denotes material conditioning and z_t is derived from the forward process described in Equation (1). After completing 15K training steps, we introduce variability in lighting conditions by integrating datasets with diverse lighting scenarios. The loss function then evolves to encompass both material and lighting factors:

$$\mathcal{L}_{\text{phys}} = \mathbb{E}_{I,C_{\text{phys}}}\left[||\epsilon - \epsilon_\theta\left(z_t, t, C_{\text{phys}}\right)||_2^2\right]. \tag{8}$$

Following an additional 15K steps, we enable semantic control by unfreezing U_{fusion}, incorporating semantic training data, and continuing the training over another 15K steps with all conditioning factors activated.

Table 1. Quantitative evaluation of material editing methods. Throughout the paper, best performances are highlighted in boldface, and ↑ (↓) means high (lower) is better. †Evaluation considers only changes in metallicity and roughness properties.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
IP2P [3]	20.85	0.69	0.31	40.80
Subias *et al.* [40]†	23.00	**0.80**	0.32	44.17
Ours	**26.01**	0.79	**0.15**	**23.38**

This training strategy ensures the model first consolidates its skills in material editing before it is exposed to varying lighting conditions, and subsequently, semantic variations. This sequential training improves disentanglement and control, allowing for precise manipulation of material properties, lighting, and semantics.

5. Experiments

5.1. Implementation Details

We initialize the parameters of PhyS-EdiT from IP2P [3] and perform image editing at a resolution of 512×512 pixels. The training is conducted over 16 days on 4 NVIDIA RTX 4090 GPUs with a total batch size of 64. We use the Adam optimizer [23] with a learning rate of 5×10^{-5}. During inference, we utilize the DDIM sampling strategy [39] with 50 sampling steps. For real images lacking ground truth environment maps, we estimate lighting conditions using an off-the-shelf lighting estimator [30]. Additional training details are provided in the supplementary material.

5.2. Evaluation Metrics

We evaluate the performance of each method using five metrics: *(i)* **FID**[17], which measures the distributional distance between generated and real images, reflecting quality and diversity. *(ii)* **SSIM**[42], which quantifies the preservation of structural information between edited images and the ground truth. *(iii)* **PSNR**[20], which evaluates the visual quality of edited regions, with higher values indicating better fidelity. *(iv)* **CLIP Score**[33], which assesses semantic alignment between edited images and text descriptions in the CLIP embedding space. *(v)* **LPIPS** [53], which measures perceptual similarity between images, focusing on human visual perception.

5.3. Quantitative Evaluation

Material editing quality. To evaluate the quality of material editing, we use a test set of 1K samples. We compare our approach with existing feed-forward-based material editing methods [3, 40], using their publicly available implementations. The results are presented in Tab. 1. Our method achieves the best PSNR, LPIPS, and FID scores, as well as the second-best SSIM.

Lighting editing quality. We evaluate the performance of our lighting editing approach using an additional 1K

Table 2. Quantitative evaluation of lighting editing methods.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DilightNet [47]	17.15	0.65	0.39	51.08
Ours	**26.96**	**0.80**	**0.13**	**21.59**

Table 3. Quantitative evaluation of semantic editing methods.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑	FID ↓
IP2P [3]	15.11	0.57	0.41	0.81	**45.14**
SD 3 [10]	14.64	0.50	0.47	0.81	52.25
Ours	**15.37**	**0.58**	**0.39**	**0.85**	47.91

test samples. We compare our results with the existing diffusion-based lighting editing method [47], which allows explicit lighting conditions as input. As shown in Tab. 2, our method achieves the highest scores across all metrics, demonstrating superior capability in lighting editing.

Semantic editing quality. We further evaluate the semantic editing capabilities of our model against state-of-the-art methods, with results shown in Tab. 3. Our method achieves the best PSNR, SSIM, LPIPS, and CLIP scores, and the second best FID.

5.4. Qualitative Evaluation

We present a visual comparison of our method against baseline approaches for lighting, material, and semantic editing, as shown in Fig. 3. For **lighting** (red columns), our method effectively adjusts lighting to match the ground truth (GT) while preserving material properties, unlike DiLightNet [47], which is not material-aware and alters the material during lighting adjustments. For **material** (green columns), our approach produces high-quality images with material properties closely matching the GT while maintaining consistent lighting, whereas Subias *et al*. [40] fails to modify attributes like albedo and transparency, and IP2P [3] often misinterprets material commands. Finally, for **semantic editing** (blue columns), our method retains strong high-level editing capabilities, achieving results comparable to or better than IP2P [3] and Stable Diffusion 3 [10].

5.5. User Study

To further assess the perceptual quality and consistency of our results, we conduct a user study involving 20 participants. The study is divided into two parts: material editing and lighting editing. In each part, participants are presented with input conditions, original images, and candidate images synthesized by different methods, including our PhyS-EdiT. They are asked to select the most visually pleasing result that best matches the input conditions (preference) and to rate the consistency of the edited images with the original content (consistency).

As shown in Tab. 4, our method receives the highest preference and consistency scores in both material and lighting editing tasks, indicating a strong user preference and better content preservation in our edited images.

Table 4. User study results showing the preference and consistency percentages for each method in material and lighting editing tasks.

Property	Metric	Method	Score (%)
Material	Preference	Ours	**68**
		IP2P [3]	18
		Subias *et al*. [40]	14
	Consistency	Ours	**61**
		IP2P [3]	11
		Subias *et al*. [40]	28
Lighting	Preference	Ours	**60**
		DiLightNet [47]	40
	Consistency	Ours	**70**
		DiLightNet [47]	30

Table 5. Ablation study results grouped by low-level and high-level editing.

Ablation	Physical Editing		Semantic Editing	
	PSNR ↑	LPIPS ↓	CLIP ↑	LPIPS ↓
Full Model	**26.39**	**0.36**	**0.85**	0.39
w/o WF	26.00	0.38	0.76	**0.35**
w/o JT	25.80	0.39	0.83	0.40

5.6. Ablation Study

We perform an ablation study to evaluate the impact of different components of our method.

Without weight freezing (WF). We train the entire network without freezing the pretrained weights. This leads to overfitting and loss of the model's generative priors, reducing the quality of the semantic edits.

Without joint training (JT). We train the material and lighting editing components separately. The lack of joint training leads to inconsistent results when both material and lighting need to be edited simultaneously.

The quantitative results of the ablation study are presented in Tab. 5, highlighting the performance of different model versions on PSNR, SSIM, LPIPS, and CLIP scores for low-level (physical) and high-level (semantic) editing tasks. The model version without weight freezing (w/o WF) achieves a slightly improved LPIPS score, as not freezing pretrained weights causes the model to forget its original semantic editing capabilities, producing nearly identical outputs regardless of the semantic editing instructions. Consequently, while the unedited image superficially resembles the edited ground truth, it fails to follow the intended instructions, leading to a notably lower CLIP score.

5.7. Discussion

Support of pixel-level editing. Our model encodes lighting and material properties as pixel-aligned condition maps, enabling precise, localized editing. For instance, in Fig. 5, we applied transparency adjustments to a plate, making it fully transparent (transparency = 1) except for the "Phys

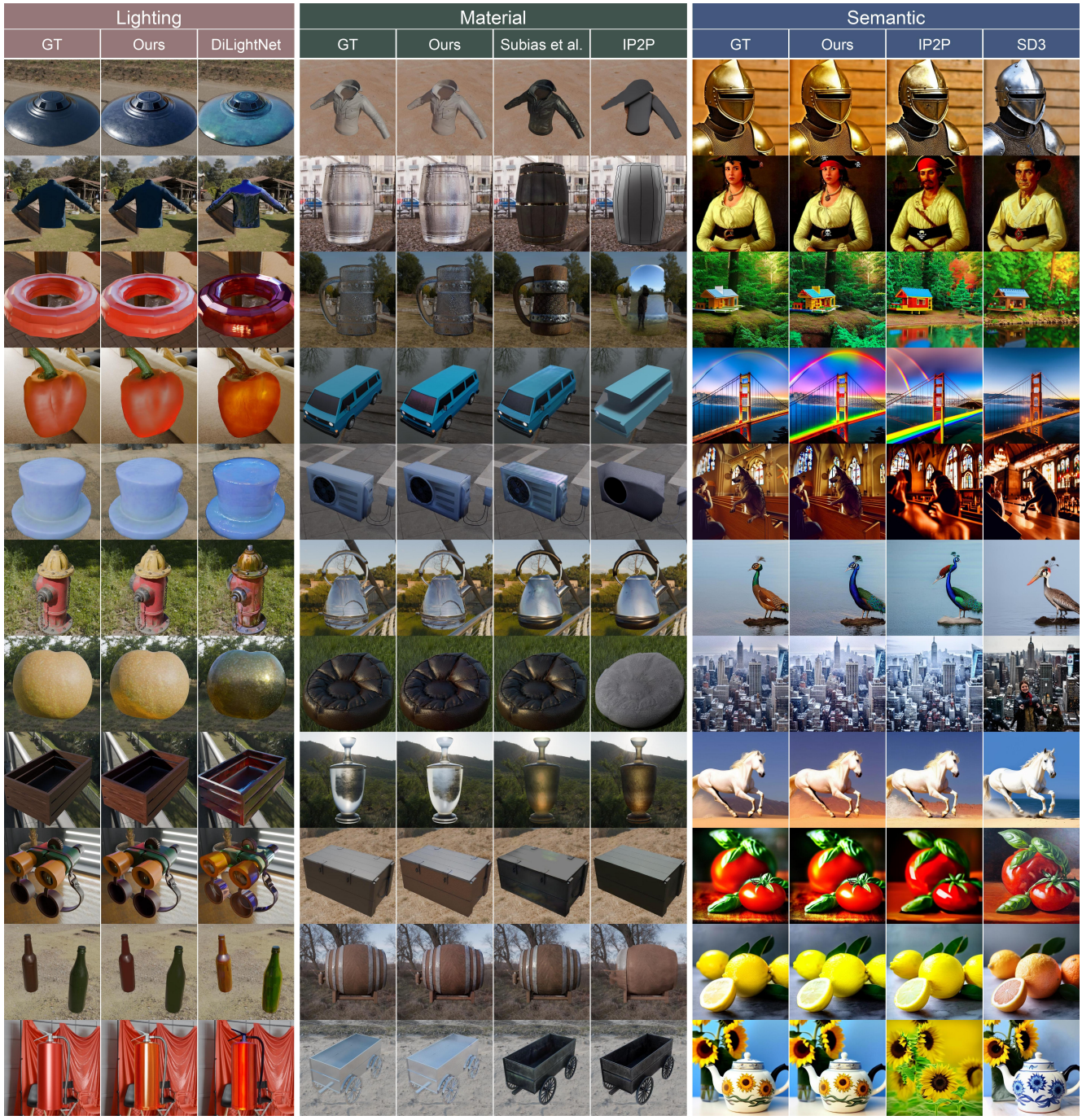

Figure 3. Visual comparison of our method with baseline approaches. For lighting, we compare with DiLightNet [47]; for material, with InstructPix2Pix [3] and Subias *et al*. [40]; and for semantics, with InstructPix2Pix [3] and Stable Diffusion 3 [10]. Our method effectively disentangles these properties, allowing for the modification of one attribute (*e.g*., lighting) while preserving others (*e.g*., material and semantics) unchanged.

Edit" text region, which remains opaque (transparency = 0).

Support of continuous property control. Our model provides smooth, continuous control over lighting and material properties using a unified map format. By feeding in maps with gradually changing values, the model generates outputs that naturally reflect these subtle, progressive adjustments. This also underscores the ability of the model to disentangle different physical conditions, allowing for independent and precise modifications. As shown in Fig. 4, row (a) demonstrates a lighting-only change, with the lighting condition visualized as a diffuse sphere. The remaining rows show combined changes in material properties along

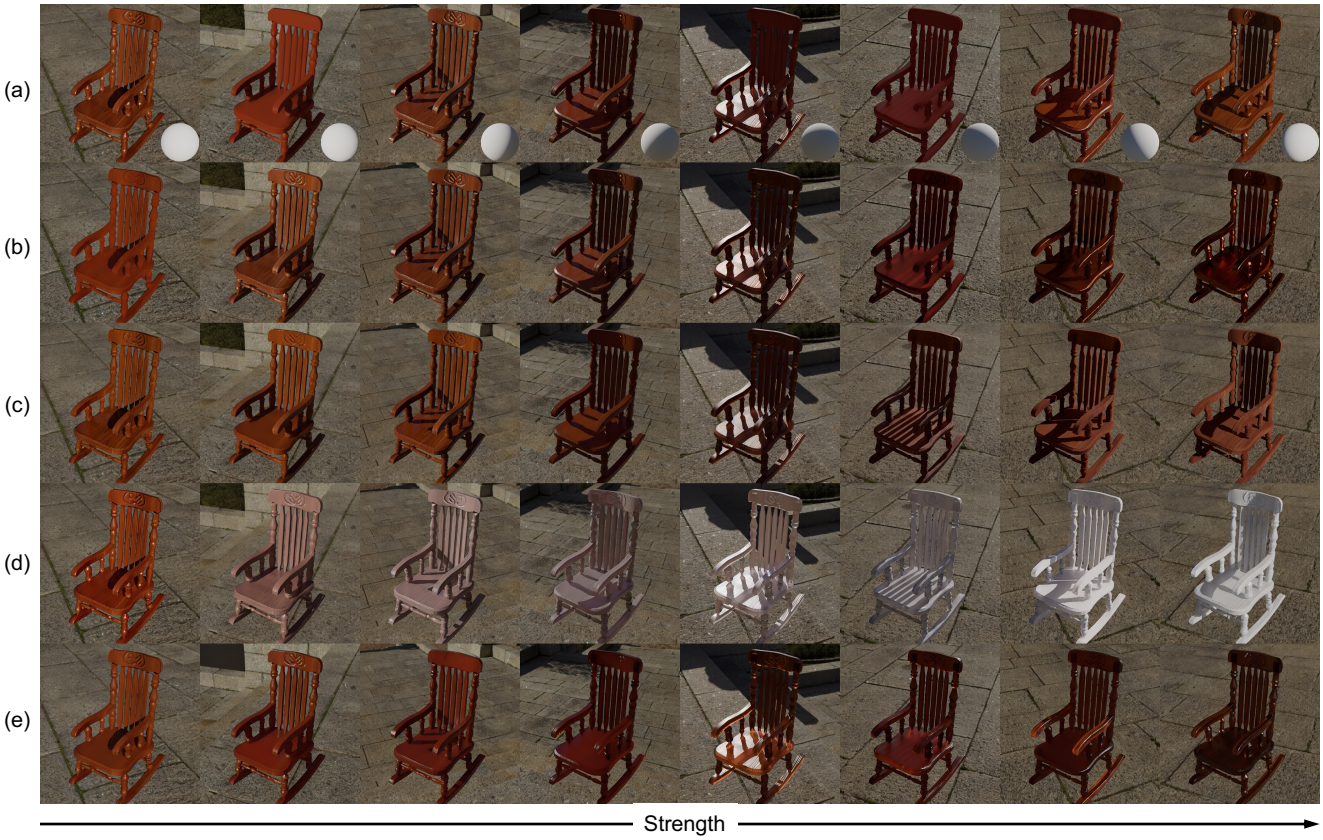

Figure 4. Demonstration of our model's capability for continuous editing across different attributes: (a) only lighting change, (b) metallicity increase, (c) roughness increase, (d) albedo increase, and (e) transparency increase. Our model produces consistent and seamless transitions.

Figure 5. Visualization of our model's capability for pixel-level editing. By applying customized condition maps to a real image, we achieve localized editing in specific properties.

with lighting: (b) increased metallicity, (c) increased roughness, (d) increased albedo, and (e) increased transparency.

Impact of multi-property control. Learning lighting and material control in conjunction improves the consistency and precision of the model. For example, by accounting for material properties, the model can adjust the lighting while keeping the material consistent. Without this material awareness, the model must rely on implicit assumptions about materials, which can lead to varied and inconsistent results across similar scenes.

Impact of Dataset Scaling. Increasing dataset scale improves performance and generalization to unseen material-lighting combinations by exposing the model to diverse lighting and material interactions. A broader dataset enables realistic editing across materials like metal and glass, while a limited dataset risks overfitting.

6. Conclusion

We introduce PhyS-EdiT, a unified image editing model that enables simultaneous control over material properties, lighting, and high-level semantics. Alongside this, we present PR-TIPS, a large-scale synthetic dataset designed to improve the disentanglement of lighting and material properties. Our approach leverages the advantages of the diffusion model, *i.e.*, disentanglement and controllability, to advance image editing grounded in physical principles. PhyS-EdiT effectively edits both synthetic and real images, supporting both localized and continuous edits.

Limitations. While our model performs precise editing and understanding of single objects, it is limited in its ability to generalize to full scene-level editing. Additionally, our model requires two forward passes for processing, which makes it slower compared to single-pass approaches. Future work could enhance scene-level capabilities through richer datasets and optimized single-pass techniques.

Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant No. 62136001 and Doubao Fund. PKU-affiliated authors thank *openbayes.com* for providing computing resources.

References

[1] Dina Bashkirova, Arijit Ray, Rupayan Mallick, Sarah Adel Bargal, Jianming Zhang, Ranjay Krishna, and Kate Saenko. Lasagna: Layered score distillation for disentangled object relighting. *arXiv preprint arXiv:2312.00833*, 2023. 2

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 4

[3] 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, 3, 4, 5, 6, 7

[4] Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. *arXiv preprint arXiv:2404.11593*, 2024. 1, 2

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

[6] Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Flexit: Towards flexible semantic image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[8] Katja Doerschner, Huseyin Boyaci, and Laurence T Maloney. Estimating the glossiness transfer function induced by illumination change and testing its transitivity. *Journal of Vision*, 2010. 2

[9] 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. 2

[10] 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, 6, 7

[11] OpenAI et al. GPT-4 technical report, 2024. 3

[12] Roland W Fleming. Visual perception of materials and their properties. *Vision research*, 2014. 2

[13] Roland W Fleming, Ron O Dror, and Edward H Adelson. Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 2003. 2

[14] Duan Gao, Guojun Chen, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. Deferred neural lighting: free-viewpoint relighting from unstructured photographs. *ACM Transactions on Graphics*, 39(6):1–15, 2020. 4

[15] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 2017. 5

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. 2, 3

[19] Yuchen Hong, Haofeng Zhong, Shuchen Weng, Jinxiu Liang, and Boxin Shi. L-DiffER: Single image reflection removal with language-based diffusion model. *European Conference on Computer Vision*, 2024. 2

[20] Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *International Conference on Pattern Recognition*, 2010. 5

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 5

[22] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[23] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 5

[25] Gihyun Kwon and Jong Chul Ye. CLIPstyler: Image style transfer with a single text condition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[26] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. ManiGAN: Text-guided image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[27] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[28] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Association for the Advancement of Artificial Intelligence*, 2024. 1

[29] Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total Relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics*, 2021. 2

[30] Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook

Khungurn, and Supasorn Suwajanakorn. Diffusionlight: Light probes for free by painting a chrome ball. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 5

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 5

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 4

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. 2

[36] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 2022. 2

[38] 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, 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 5

[40] J Daniel Subias and Manuel Lagunas. In-the-wild material appearance editing using perceptual attributes. In *Computer Graphics Forum*, 2023. 1, 2, 5, 6, 7

[41] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. SinSR: diffusion-based image super-resolution in a single step. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022. 3

[44] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. CT^2: Colorization transformer via color tokens. In *European Conference on Computer Vision*, 2022. 2

[45] Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, Boxin Shi, et al. L-CAD: Language-based colorization with any-level descriptions using diffusion priors. In *Advances in Neural Information Processing Systems*, 2024. 2

[46] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2

[47] 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, 2, 4, 6, 7

[48] Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. RGB↔X: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *ACM SIGGRAPH Conference Papers*, 2024. 1, 2, 4

[49] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *International Conference on Computer Vision*, 2017. 2

[50] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. In *ACM International Conference on Multimedia*, 2020. 2

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, 2023. 1, 4

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *International Conference on Learning Representations*, 2025. 1, 2

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 5

[54] Haofeng Zhong, Yuchen Hong, Shuchen Weng, Jinxiu Liang, and Boxin Shi. Language-guided image reflection separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

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.

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.

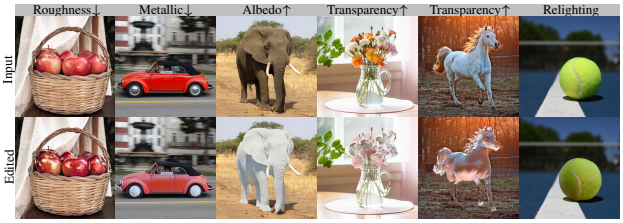

Figure A. Real-image results.

D.2. Retraining IP2P

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

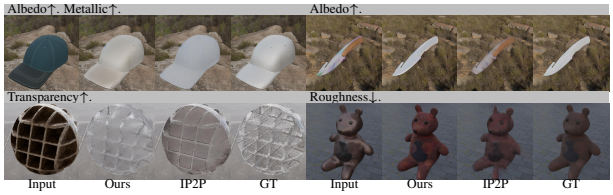

Figure B. Comparison to retrained IP2P.

*Corresponding author.

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.

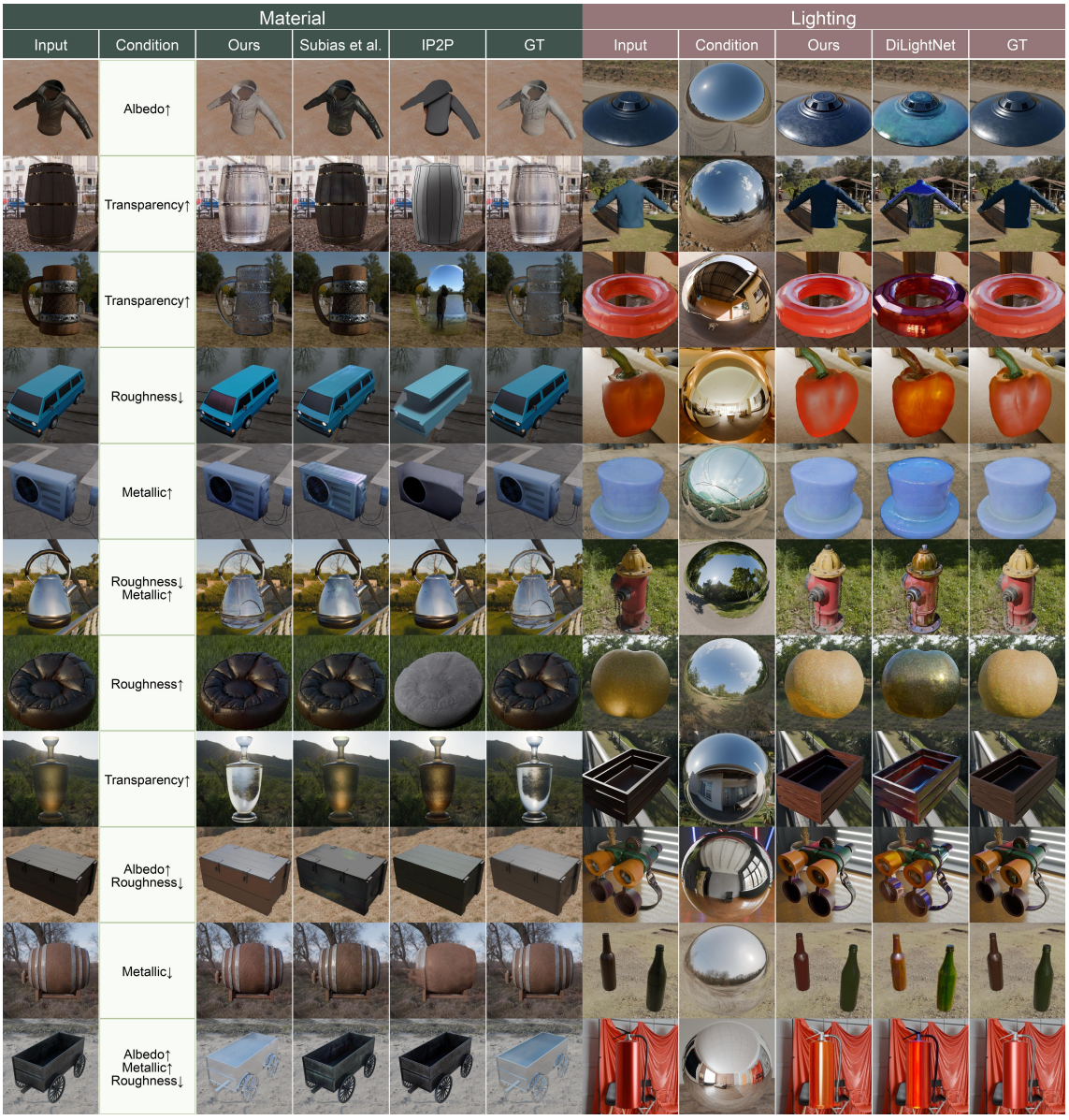

Figure E. The complete visualization for material editing and lighting editing, including input, condition, output, and ground truth.

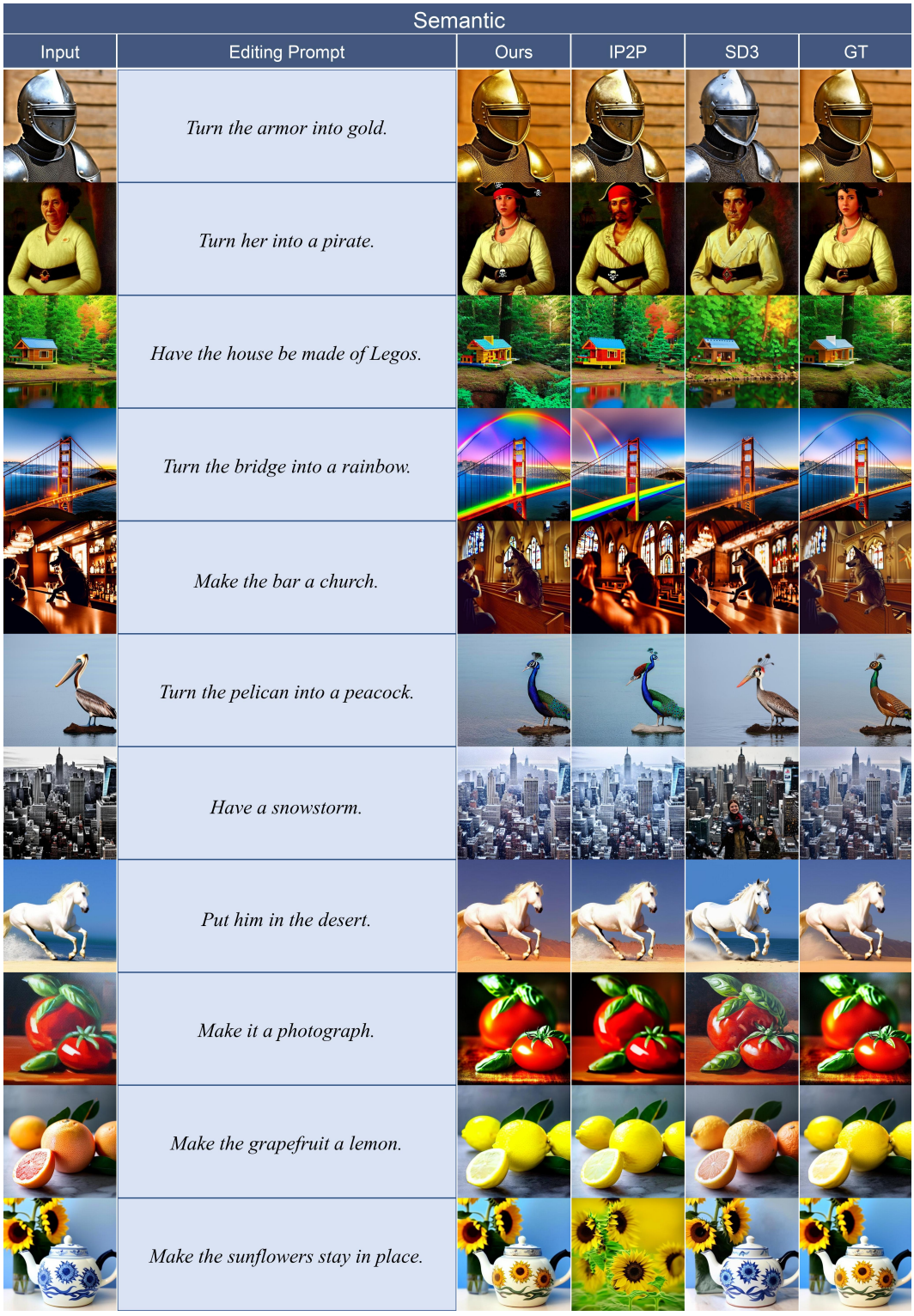

Figure F. The complete visualization for semantic editing, including input, condition, output, and ground truth.

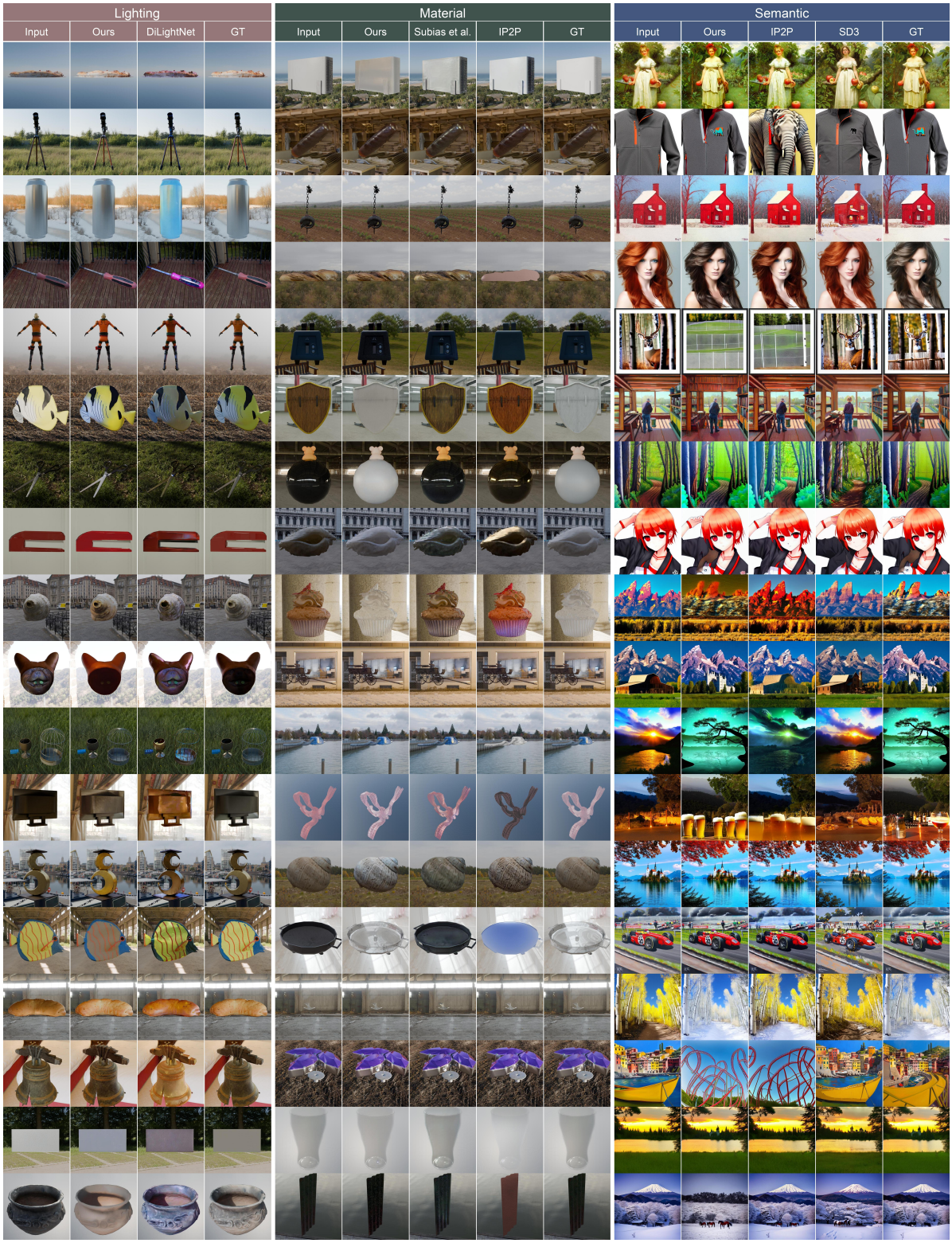

Figure G. Additional comparison results for material, lighting, and semantic editing (specific conditions omitted for clarity).

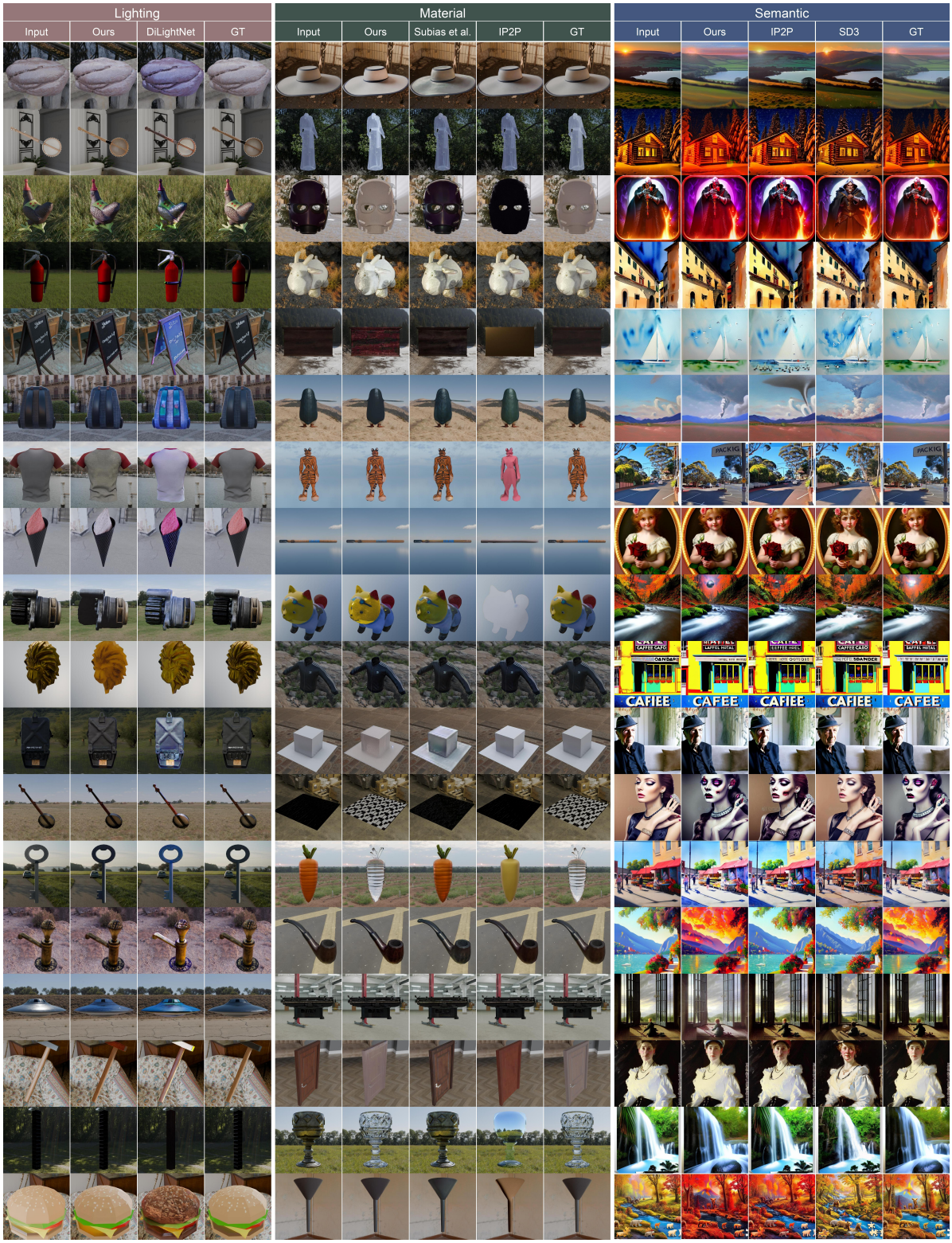

Figure H. Additional comparison results for material, lighting, and semantic editing (specific conditions omitted for clarity).