Eurographics Symposium on Rendering 2025
B. Wang and A. Wilkie
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 4

SUPPLEMENTAL MATERIAL - Artist-Inator: Text-based, Gloss-aware Non-photorealistic Stylization

J. Daniel Subias[1]*, Saul Daniel-Soriano[1], Diego Gutierrez[1] & Ana Serrano[1]

[1]Universidad de Zaragoza, I3A, Spain
*dsubias@unizar.es

The supplemental material of this paper includes:

- (S1) Additional images of paintings created by the other three artists not shown in the main paper.
- (S2) Additional details of the procedure of our user study.
- (S3) Additional information and details of our datasets.
- (S4) Additional details of our user study procedure for assessing the fidelity of the generated painterly depictions with respect to the input text descriptions.
- (S5) Additional results and comparisons of our method.

S1. Paintings: Additional Details

Figures 1, 2 and 3 show the paintings created by the other three artists (not shown in the main paper) of the four spheres with different gloss levels by varying the roughness parameter r of the *Disney's Principled BSDF* [BS12, Bur15] within the set of values $\{0.0, 0.24, 0.47, 0.7\}$ with a fixed albedo 0.18.

S2. User Study: Additional Details

During the experiment, the images were shown randomly at a resolution of 512x512 px. , without repeating the stimuli. Figure 4 shows a screenshot of the perceptual study, as seen by the participants. The stimulus is shown on the left part of the screen while the list of ratings (from 1 to 7) is shown on the right.

S3. Datasets: Additional Details

S3.1. Training Dataset

Our training dataset comprises single-object scenes with 41 different geometries. Figure 5 shows the geometries present in our training dataset and their corresponding text labels to derive our automatically-computed text descriptions. To define the text labels for all 23 colors in our training dataset, we group the colors in eight clusters with the same text label, as is shown in Figure 6. The painterly depictions generated using the paintings created by the four artists are associated with the textual label *"gray"* for color. To increase the color variance, we apply a random shift on the saturation channel on the HSV color space between -20º and 20º, when colorizing the sphere paintings using the colorize function available in *GIMP* [Gim]. The four illuminations were generated by rotating an area light around the object in different positions (see Figure 7, left). We further sample the objects' surface by generating two additional random rotations along the axis y between -90º and 90º for each combination of geometry, illumination, and gloss level (see Figure 7, right). As materials, we generate seven gloss levels using the *Disney's Principled BSDF* [BS12, Bur15], for which we vary the value of the roughness parameter r within the set of values $\{0.0, 0.12, 0.24, 0.35, 0.47, 0.58, 0.7\}$. We styled each of the renders using StyLit [FJL*16], taking as reference both the spheres painted by each of the four artists and their color versions. The renders with a roughness value $r \in \{0.0, 0.7\}$, were stylized with the paintings of the spheres with the same roughness value r. The renders with a roughness value $r \in \{0.12, 0.24, 0.35\}$ were stylized using the paintings of the sphere with the roughness value $r = 0.24$. Similarly, the renders with a roughness value $r \in \{0.47, 0.58\}$ were stylized using the paintings of the sphere with the roughness value $r = 0.47$.

S3.2. Evaluation Dataset

Our test dataset includes 11 new geometries (see Figure 8); and the same colors and illuminations present in the training dataset (see Figures 6 and 7). Following the same procedure as for the training dataset, we generate seven gloss levels using the *Disney's Principled BSDF* [BS12, Bur15], for which we vary the value of the roughness parameter r within the set of values $\{0.0, 0.12, 0.24, 0.35, 0.47, 0.58, 0.7\}$. We styled each of the renders using StyLit [FJL*16], taking as reference both the spheres painted by each of the four artists and their color versions. The renders with a roughness value $r \in \{0.0, 0.7\}$, were stylized with the paintings of the spheres with the same roughness value r. The renders with a roughness value $r \in \{0.12, 0.24, 0.35\}$ were stylized using the paintings of the sphere with the roughness value $r = 0.24$. Similarly, the renders with a roughness value $r \in \{0.47, 0.58\}$ were stylized using the paintings of the sphere with the roughness value $r = 0.47$.

S4. User Study for Evaluation: Additional Details

During the experiment, the images were shown without repeating the stimuli. Figure 9 shows a screenshot of the user study, as seen

by the participants. The grayscale painting of a reference sphere created by one of the artists is shown on the top, while the stimuli, both text description and strip of painterly depictions, are shown in the middle. The users have to rank the painterly depictions in the strip using a grid of several options at the bottom, where columns and rows refer to painterly depictions and ranking places (from first to fifth), respectively.

To generate the stimulus, we select some condition images (edge maps, clip arts, and hand-drawn sketches) from our evaluation dataset and generate new text descriptions following the template *"A [GL] [G] in [C] [S]"*. Where [*GL*] represents the gloss level, [*G*] refers to the geometry (shape) of the object, [*C*] refers to the color, and [*S*] corresponds to the hand-drawn artistic style (i.e., oil painting, watercolor, ink pen, charcoal, and soft crayon). In total, 20 painterly depictions per diffusion model were generated, 10 in matte gloss level and 10 in glossy one, with approximately four painterly depictions per style and random colors from the list: green, blue, orange, brown and red; except for those in charcoal style, which were generated in gray.

S5. Additional Results

We provide additional results of our framework: for edge maps, sketches and clip arts (see Figure 10); extra comparisons with StyLit [FJL*16] using different inputs (see Figure 11) and only reducing it to an input more similar to ours (Figure 12); and comparisons with other state-of-the-art methods: of ControlNet [ZRA23], T2I-Adapter [MWX*24], ControlNet++ [LYK*24], and a pre-trained version of Stable Diffusion XL [PEL*23] conditioned by ControlNet but not fine-tuned on our dataset (see Figure 13).

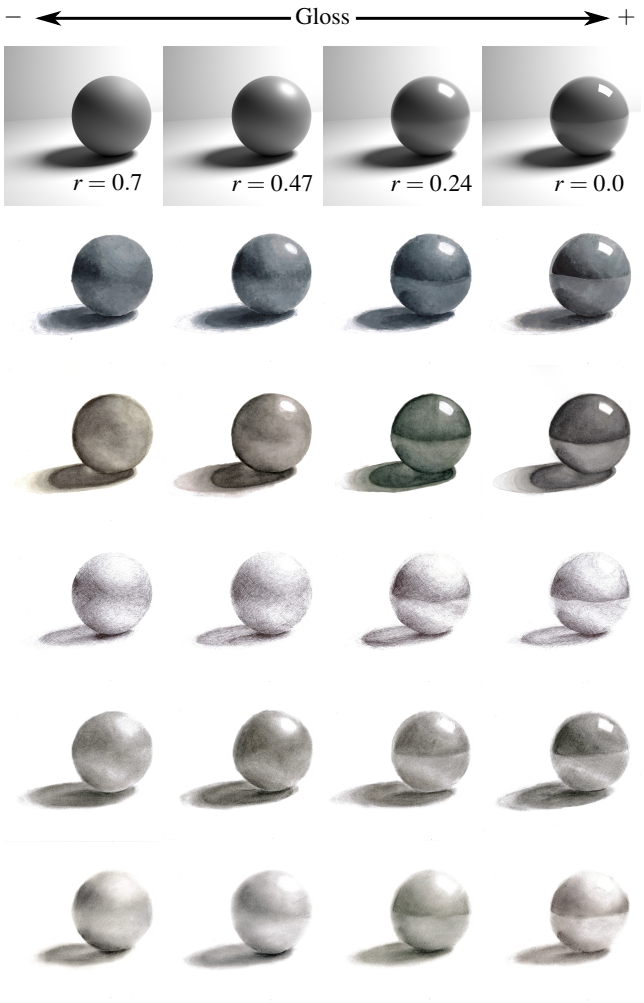

Figure 1: *The first row shows the four photorealistic reference renders with varying gloss levels, used by the artists as guides, the numbers on the bottom right corners are the roughness values r of the Disney's Principled BSDF [BS12, Bur15] used during rendering. The subsequent rows present the corresponding paintings created by the second artist for each of the five hand-drawn artistic styles featured in our dataset: oil painting, watercolor, ink pen, charcoal, and soft crayon (from second row onward).*

References

[BS12] BURLEY B., STUDIOS W. D. A.: Physically-based shading at disney. In *ACM SIGGRAPH* (2012), vol. 2012, vol. 2012, pp. 1–7.

[Bur15] BURLEY B.: Extending the disney brdf to a bsdf with integrated subsurface scattering. *SIGGRAPH Course: Physically Based Shading in Theory and Practice. ACM, New York, NY 19* (2015).

[FJL*16] FIŠER J., JAMRIŠKA O., LUKÁČ M., SHECHTMAN E., ASENTE P., LU J., SÝKORA D.: StyLit: Illumination-guided example-based stylization of 3D renderings. *ACM Transactions on Graphics 35*, 4 (2016).

[Gim] GIMP: Gimp. URL: https://www.gimp.org.

[LYK*24] LI M., YANG T., KUANG H., WU J., WANG Z., XIAO X., CHEN C.: Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision (ECCV)* (2024).

[MWX*24] MOU C., WANG X., XIE L., WU Y., ZHANG J., QI Z., SHAN Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2024), vol. 38, pp. 4296–4304.

[PEL*23] PODELL D., ENGLISH Z., LACEY K., BLATTMANN A., DOCKHORN T., MÜLLER J., PENNA J., ROMBACH R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[ZRA23] ZHANG L., RAO A., AGRAWALA M.: Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 3836–3847.

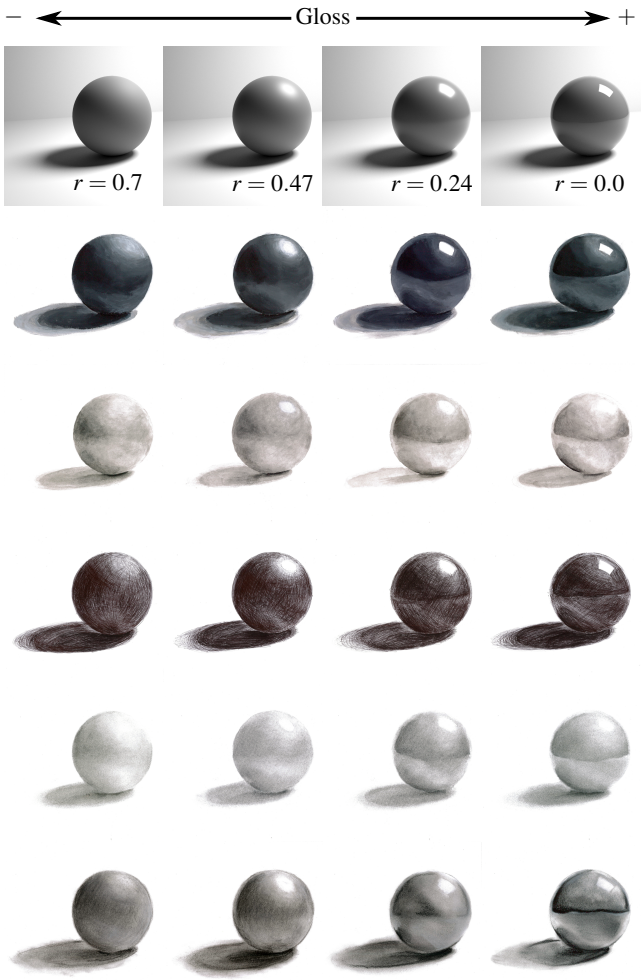

Figure 2: *The first row shows the four photorealistic reference renders with varying gloss levels, used by the artists as guides, the numbers on the bottom right corners are the roughness values r of the Disney's Principled BSDF [BS12, Bur15] used during rendering. The subsequent rows present the corresponding paintings created by the third artist for each of the five hand-drawn artistic styles featured in our dataset: oil painting, watercolor, ink pen, charcoal, and soft crayon (from second row onward).*

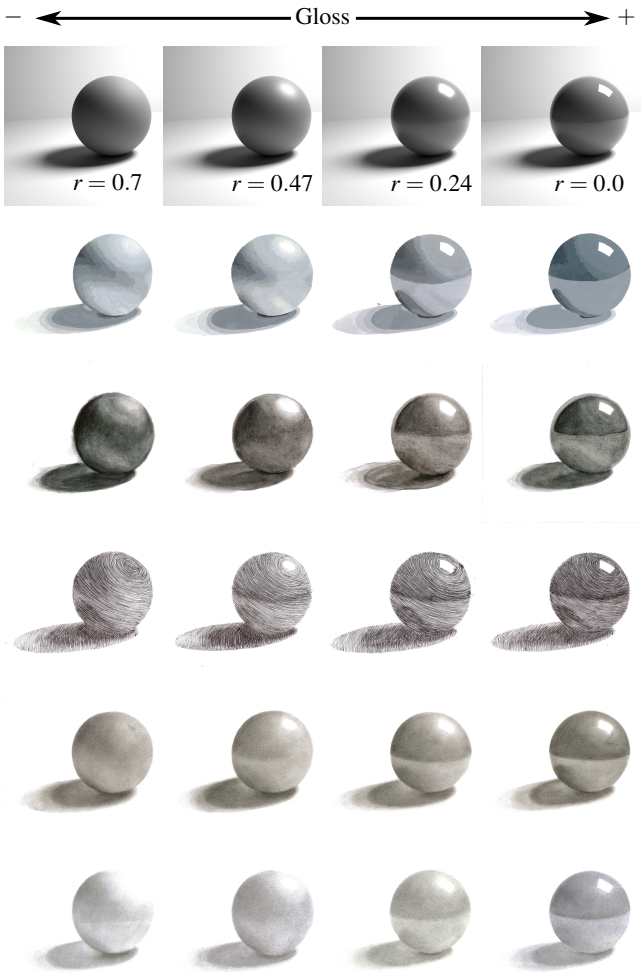

Figure 3: *The first row shows the four photorealistic reference renders with varying gloss levels, used by the artists as guides, the numbers on the bottom right corners are the roughness values r of the Disney's Principled BSDF [BS12, Bur15] used during rendering. The subsequent rows present the corresponding paintings created by the fourth artist for each of the five hand-drawn artistic styles featured in our dataset: oil painting, watercolor, ink pen, charcoal, and soft crayon (from second row onward).*

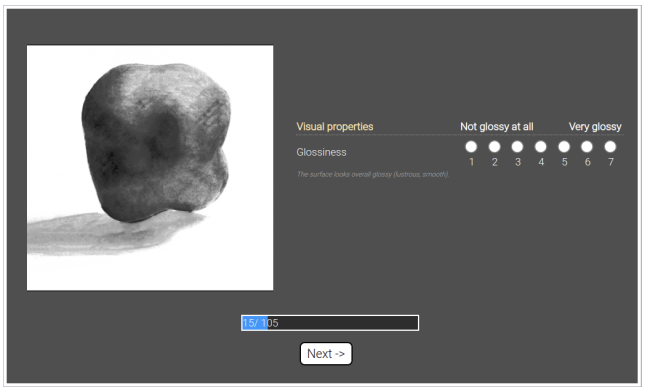

Figure 4: *Screenshot of the user study as seen by the annotators. Stimuli is shown on the left, the annotators have to select a rating for the gloss level on the right.*

Figure 5: *The 41 geometries present in our training dataset under one of the four illuminations and with a gloss gray material. The words in the lower left boxes refer to the text label associated with each geometry.*

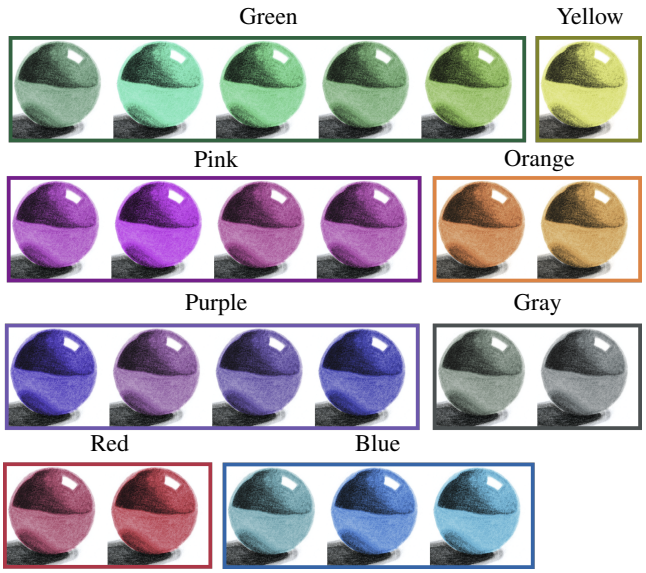

Figure 6: *All 23 colors present in our training and evaluation datasets in soft crayon style. Colors in the same color box have the same text label.*

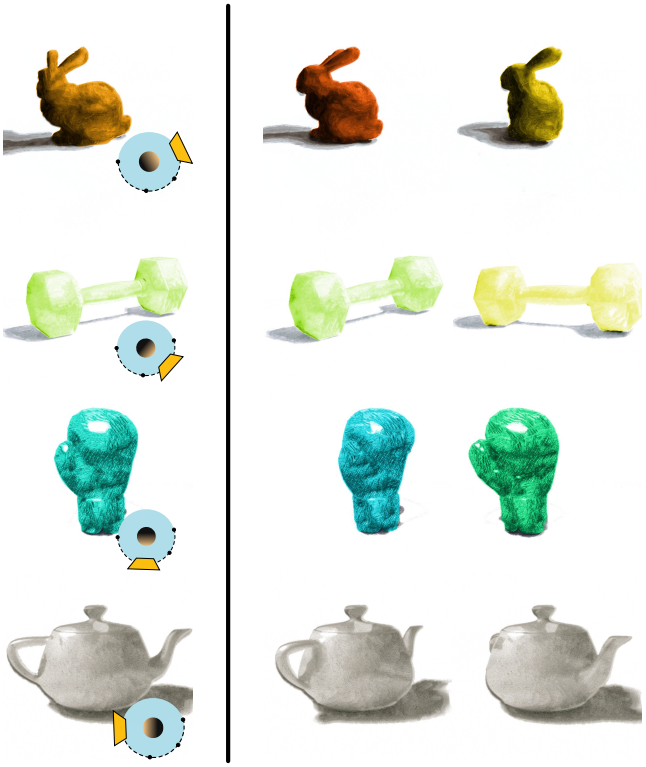

Figure 7: *Examples of the four area-light illuminations (depicted in the bottom left corners of the images in the left) present in our dataset for the "bunny", "dumbbell", "boxing glove" and "teapot" geometries with different randomly-colored styles (left). Two random rotations about the y axis between -90º and 90º (right).*

Figure 8: *The 11 geometries present in our evaluation dataset under one of the four illuminations and with gray glossy material.*

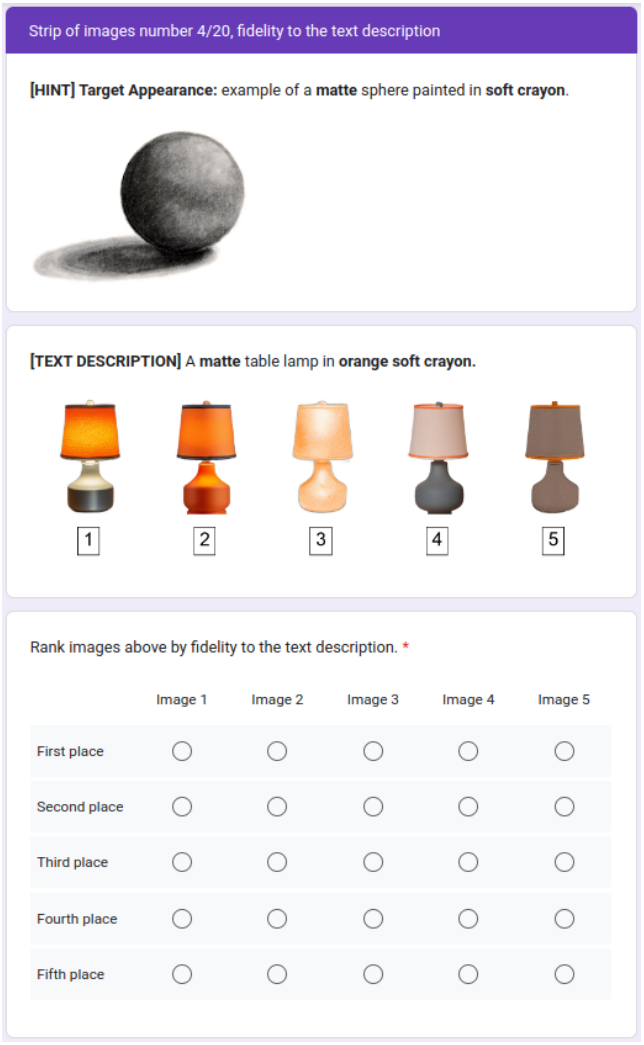

Figure 9: *A screenshot of the user study as seen by the users. To help participants understand the intended artistic traits, a grayscale painting of a reference sphere created by one of the artists is shown on the top, while the stimuli is in the center. The user has to rank the painterly depictions in the strip using the grid of several options at the bottom, without the possibility of choosing more than one option per column (assigning two positions to the same image).*

Figure 10: *Results of painterly depictions from different types of condition images: edge maps, clip arts, and hand-drawn sketches; varying the style with a fixed color (except for charcoal style). We can observe how our framework synthesizes consistent painterly depictions for all styles in our dataset, according to the input prompt, while following the semantics of the input prompt.*

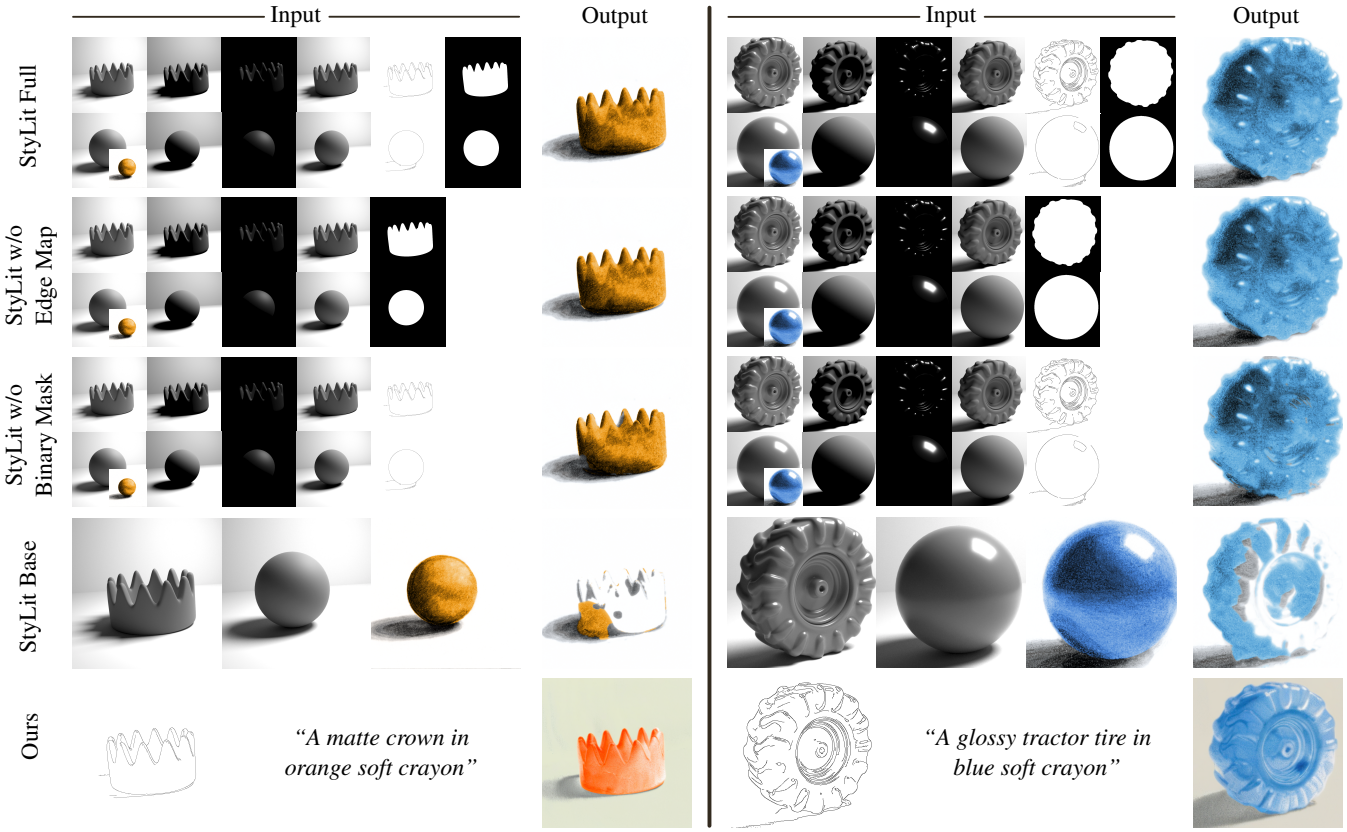

Figure 11: *Two additional comparisons of painterly depictions generated using StyLit [FJL*16] with the full input set (StyLit Full, first row), without using the edge map (StyLit w/o Edge Map, second row), without using the binary mask (StyLit w/o Binary Mask, third row), and reducing it to an input more similar to ours (StyLit Base, fourth row); and our proposed method (fifth row). In contrast, our framework input only includes an edge map and a prompt. Our framework yields more accurate results with just an edge map (shown) plus a text prompt describing the desired output, while StyLit is strongly dependent on the LPEs.*

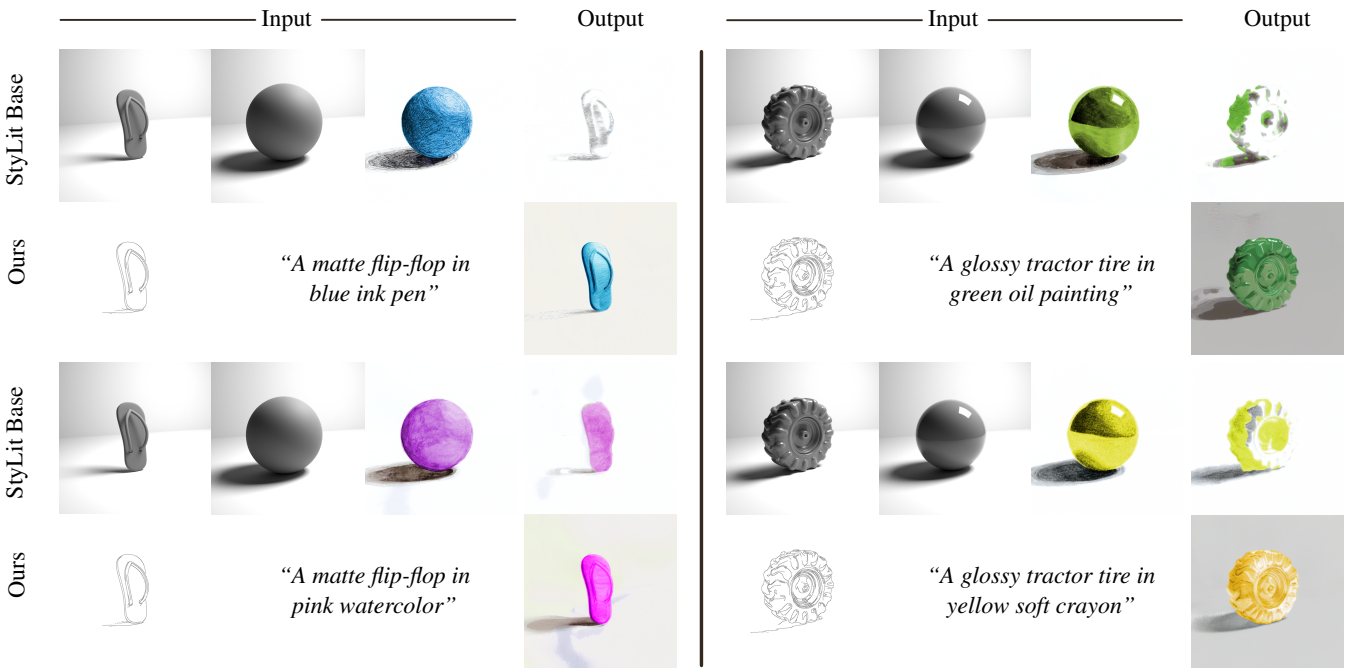

Figure 12: *Additional comparisons of painterly depictions generated using StyLit [FJL*16] reducing it to an input more similar to ours (StyLit Base); and our proposed method. In contrast, our framework input only includes an edge map and a prompt. Our framework yields more accurate results with just an edge map (shown) plus a text prompt describing the desired output, while StyLit is strongly dependent on the LPEs.*

Figure 13: *Qualitative comparison to previous methods: ControlNet (v1.1) [ZRA23], T2I-Adapter [MWX*24], ControlNet++ [LYK*24], and pre-trained Stable Diffusion XL [PEL*23] conditioned with ControlNet but not trained on our dataset (ControlNet SDXL); and our method. We can observe how our method leads to much better painterly depictions inferring the semantics from the input condition image, while depicting a visually compelling appearance according to the input prompt.*