DOI: 10.1111/cgf.70182
Eurographics Symposium on Rendering 2025
B. Wang and A. Wilkie
(Guest Editors)

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

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

Figure 1: *Top row: Example painterly depictions from our dataset. We build a non-photorealistic dataset made up of 1,336,272 painterly depictions of a large variety of objects in several colors and hand-drawn artistic styles (i.e., oil painting, watercolor, ink pen, charcoal and soft crayon), including automatically-computed descriptions of their appearance. Bottom row: We then leverage our dataset to train a framework based on Stable Diffusion XL that enables intuitive synthesis of novel painterly depictions described with a simple text prompt. In contrast with other methods that require a complex input, our framework works with simple edge maps, hand-drawn sketches, or clip arts. The example shows the input clip art and the results of the prompt: "A matte car in [red oil painting, red watercolor, red ink pen, gray charcoal, and red soft crayon]".*

Abstract

Large diffusion models have made a remarkable leap synthesizing high-quality artistic images from text descriptions. However, these powerful pre-trained models still lack control to guide key material appearance properties, such as gloss. In this work, we present a threefold contribution: (1) we analyze how gloss is perceived across different artistic styles (i.e., oil painting, watercolor, ink pen, charcoal, and soft crayon); (2) we leverage our findings to create a dataset with 1,336,272 stylized images of many different geometries in all five styles, including automatically-computed text descriptions of their appearance (e.g., "A glossy bunny hand painted with an orange soft crayon"); and (3) we train ControlNet to condition Stable Diffusion XL synthesizing novel painterly depictions of new objects, using simple inputs such as edge maps, hand-drawn sketches, or clip arts. Compared to previous approaches, our framework yields more accurate results despite the simplified input, as we show both quantitative and qualitatively.

CCS Concepts

• ***Computing methodologies*** → ***Non-photorealistic rendering; Image processing; Perception;***

1. Introduction

Non-photorealistic rendering (NPR) aims to create images that emphasize aesthetics or convey information in a more stylized or artis-

tic manner. Since the formulation of image analogies [HJO*01], methods based on patch matching have been proposed to stylize images of 3D models. However, these methods require the help of additional input, such as a stylized image of a sphere with the desired style, and several rendered maps encoding normals, direct illumination, specular highlights, and first and second diffuse light bounces [FJL*16, SJT*19]. The advent of large diffusion models [RBL*21, PEL*23, RDN*22] has simplified this process by generating artistic images from simple text prompts. However, these models lack control over key features like gloss, which is the focal point of this work. Gloss is a fundamental aspect of surface reflectance, essential for material recognition [CK15, MKA12, GOS*10, SCW*21] and key in fields beyond computer graphics, like experimental psychology [And11, Fle17], or fabrication [CJP*23, CPBD23].

Bousseau et al. [BOD*13] analyzed how gloss is perceived across three different computer-generated artistic styles: painterly rendering, cartoon rendering, and Gaussian blur. The authors conclude that shiny materials appear more diffuse when depicted in painterly and cartoon styles, while diffuse materials appear shinier in cartoon styles. However, it is not clear if and how these conclusions extrapolate to other hand-drawn styles beyond those three. Recently, Zuijlen et al. [vZPW20] analyzed how humans perceive high-level perceptual attributes (e.g., gloss, roughness, or hardness) in paintings, concluding that material perception operates independently of the representation medium (i.e., paintings and photos). Close to this work, Delanoy et al. [DSMG21] yielded similar conclusions, finding that perception of gloss in painterly depictions is linked to similar visual cues than in photorealistic stimuli.

In this work, we first analyze how gloss is perceived across five different hand-drawn artistic styles: oil painting, watercolor, ink pen, charcoal, and soft crayon. We build a large non-photorealistic dataset made up of 1,336,272 stylized versions of different objects including all five styles (Figure 1, first row), annotated with automatically-computed text descriptions of their appearance based on the findings of our user study. Finally, we use our non-photorealistic dataset to train a framework based on Stable Diffusion XL [PEL*23] to synthesize novel painterly depictions with a hand-drawn artistic appearance from three kinds of simple input images: edge maps, hand-drawn sketches, and clip arts (Figure 1, second row). Our results compare favorably against other state-of-the-art models both qualitatively and quantitatively, despite requiring a much simpler input.

Our dataset and model, as well as the training and evaluation code, are available at https://graphics.unizar.es/projects/artist-inator_2025/

2. Related Work

2.1. Non-photorealistic Rendering

NPR has the potential to create artistic representations of synthetic scenes. We provide here a brief cross-section of different NPR techniques and refer to the survey of Kumar et al. [KPNM19] for a more comprehensive review.

Over the years, many approaches have been proposed [KK87, DS02] including styles such as painterly [Hae90, Mei96, Lit97, Her98, HE04, ZZXZ09] or cartoonish [WFGS07]. The work of Paul Haeberli [Hae90] introduced the concept of style transfer, aiming to convert a photo or rendered image into a stylized image. Following this line of work, numerous techniques have been proposed, using procedural techniques [BLV*10, BKTS06], image filtering [WKO12, LXJ12], or composition of exemplar strokes [SWHS97, ZZ11].

Building on the patch-based formulation of image analogies introduced by Hertzmann et al. [HJO*01], Fischer et al. presented StyLit [FJL*16]. This method offers an efficient and controllable stylization of a photorealistic render of a single object, starting from a stylized image of a lit sphere [SMGG01] with similar materials, lighting conditions and spatial position. It was later extended to face stylization in videos [FJS*17] and real time [SJT*19]. With the emergence of deep learning, the style transfer research domain has tended towards approaches based on convolutional neural networks (CNNs). Gatys et al. [GEB15] introduced the idea of representing image styles as high-level features extracted from pre-trained CNNs, which paved the way for more deep-learning-based methods for image and video stylization [CS16, JAFF16, KLA19, CLY*17, DTD*21]. Close to our work, Futschik et al. [FCC*19] generated a large dataset of image pairs (source images and their stylized counterparts) and then trained a custom variant of U-Net [RFB15], focusing on real-time face stylization. While producing excellent results, CNN-based methods lack user control, since the style transfer process is based on learned statistics of color patterns, and struggle to preserve high-frequencies and low-level details [LDX*19, FCC*19, DLGM22, SL23], leading to incoherent results for the user.

With the explosion of diffusion models, the classical image analogies formulation [HJO*01] was extended to work with unaligned images and changes in higher-level semantics [vLv*23]. However, the generative nature of diffusion models accentuates this lack of control over the final image. To avoid this, ControlNet [ZRA23, MWX*24, LYK*24, BBSM*25] allows conditioning diffusion models with additional information of the desired scene (e.g., edge maps, depth maps or camera parameters), to generate stylized content from a prompt. However, despite the existence of large datasets to condition diffusion models for sketch-to-image tasks [QZY*23, ZRA23], the lack of data specialized in specific hand-drawn artistic styles makes it difficult for diffusion models to learn to simulate visual features depicted by artists when painting. In our work, we address this by leveraging StyLit [FJL*16] to generate a large dataset of painterly depictions from human paintings in several hand-drawn artistic styles (i.e., oil painting, watercolor, ink pen, charcoal, and soft crayon). We demonstrate the applicability of our dataset by training ControlNet to condition Stable Diffusion XL to synthesize paintings given a semantic condition image (e.g., edge map, hand-drawn sketch, or clip art) of a single object. We also show that despite resorting to simpler input, our diffusion model obtains results that better preserve the semantics in comparison with StyLit.

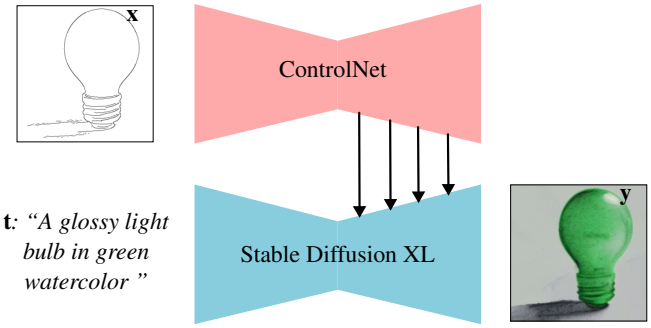


Figure 2: *Overview of our framework. Given an input condition image* **x** *(shown: an edge map) and a text prompt* **t***, our framework synthesizes a painterly depiction* **y** *of the object. The zero convolution feature maps of ControlNet [ZRA23] are added to the residual connections between the encoder and the decoder, of Stable Diffusion XL [PEL*23] (black arrows).*

2.2. Text Description of Visual Attributes

Using natural language descriptions, it is possible to provide additional semantic or comparative information which could not be obtained from a simple numerical rating [FEHF09]. In the work of Bhushan et al. [BRL97], the authors investigate the lexicon used by humans to describe visual textures. During the experiments, they discovered that a simple 98-word lexicon could describe up to 82% of their experimental data. In the same line of work, Cimpoi et al. [CMK*14] show that a simple lexicon of 47 texture words suffices to describe natural patterns. This work was later expanded by Wu et al. [WTM20] to include natural language descriptions. Recently, Deschaintre et al. [DGVG*23] collect and analyze a dataset that links free-text descriptions of textiles. They identify a compact lexicon of a set of attributes (e.g., color, pattern, touch, etc.) that are relevant when describing fabrics. Close to our work, Butt et al. [BWVCW24] condition Stable Diffusion [RBL*21] to learn to synthesize specific colors from a dataset of images of objects with simple shapes linked with automatically generated text descriptions. Inspired by this, we have collected a dataset made up of high-quality painterly depictions in different artistic styles (i.e., ink pen, charcoal, soft crayon, watercolor, and oil painting) linked with automatically-computed text descriptions of their appearance.

3. Goal and Overview

ControlNet [ZRA23] has had a major impact on conditioning large pre-trained text-to-image models, such as Stable Diffusion [RBL*21, PEL*23]. Our goal is to synthesize a painterly depiction **y** of an object, given a condition image **x** (an edge map, a hand-drawn sketch or a clip art). The information of the desired artistic appearance is provided in the input prompt **t**, which encodes our target visual attributes: gloss level, style, and color (for instance: *"A matte light bulb in green watercolor"*).

In particular, ControlNet is a U-Net [RFB15] with an encoder and a decoder connected via skip connections. The encoder part of the U-Net is a trainable copy of the encoder of Stable Diffusion XL [PEL*23], which is concatenated with zero convolution

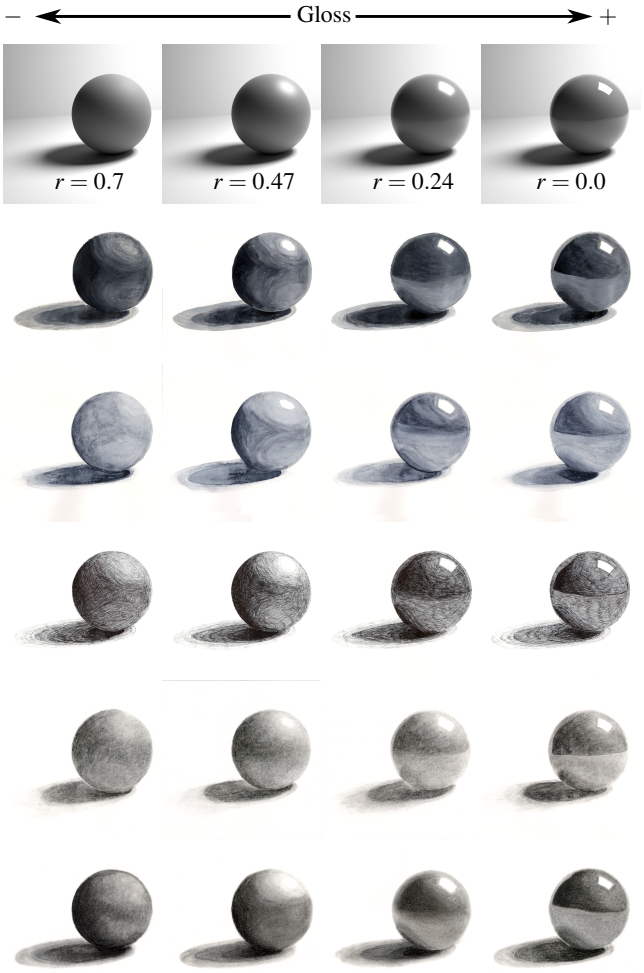


Figure 3: *The first row shows the four photorealistic reference renders with varying gloss levels used by the artists as guides. The numbers on the bottom right corners are the roughness values r of the Disney's Principled BSDF [BS12, Bur15] used during rendering. The subsequent rows present the corresponding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

layers, whose feature maps are added to the residual connections, between the encoder and the decoder, of Stable Diffusion XL (see Figure 2). During training, the ControlNet module learns the conditional control from the semantics of the condition image **x**, while Stable Diffusion XL retains the knowledge learned from billions of images in its current version (for further details see the original work of Podell et al. [PEL*23]). To train our framework, we begin by gathering hand-painted references of five hand-drawn artistic styles: oil painting, watercolor, ink pen, charcoal, and soft crayon (Section 4). Next, we conduct a user study to analyze gloss perception across all these hand-drawn artistic styles, assessing whether the perception of gloss in our painterly depictions aligns with that

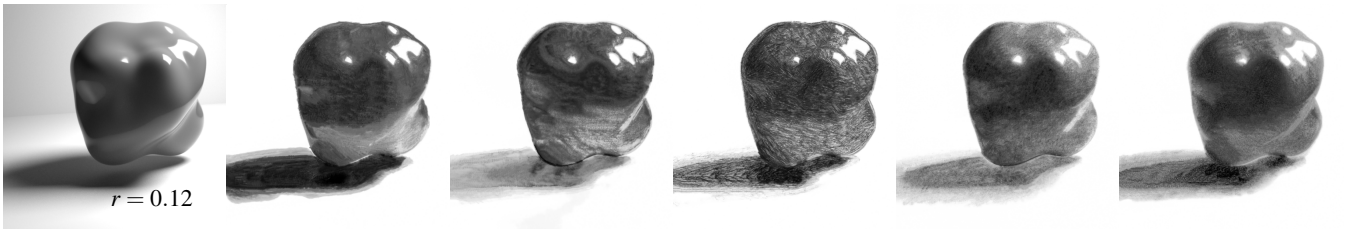


Figure 4: *Examples of stimuli shown to the participants in the study. A photorealistic render of the blob geometry, where the number on the bottom right corner is the roughness value r of the Disney's Principled BSDF [BS12, Bur15] used during rendering (first column). The subsequent columns present the corresponding stylized versions of the render generated with StyLit [FJL*16] on the five hand-drawn styles: oil painting, watercolor, ink pen, charcoal, and soft crayon; using the paintings created by one of the artists (from second column onward).*

in their photorealistic counterparts (Section 5). Finally, we (1) use the hand-painted references to generate a dataset of over 1.3 million painterly depictions using StyLit [FJL*16] and (2) leverage our findings on gloss perception to annotate these painterly depictions with automatically-computed text descriptions of their appearance (Section 6).

4. Gathering Hand-Painted References

We focus on five hand-drawn artistic styles that present notable differences in their aesthetic qualities: oil painting and watercolor for the presence of brushstrokes; ink pen for the presence of strokes; and charcoal and soft crayon for their stippled texture.

To gather high-quality references for these five styles, we collaborated with four local artists with formal artistic training, who were economically compensated for their time and effort. We provided them with reference renders of a photorealistic gray sphere, which was rendered using Mitsuba 3 [JSR*22]. We focus on monochromatic objects since they allow us to study the effects of style and gloss, isolating color side effects. This choice aligns with existing works in both photorealistic [SCW*21, GVSS*24, AKLM18] and non-photorealistic rendering [BOD*13, DSMG21]. Since we were interested in how different materials are depicted, we rendered four spheres by varying gloss.We focus on gloss since glossiness is arguably one of the most important material appearance attributes [And11, MKA12, CK15, Fle17, SCW*21], alongside with color. The different gloss variations were generated 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 (see Figure 3, first row). The spheres were illuminated by an area light, placed on the top right corner of the scene.Each artist was asked to depict the four sphere renders in the five selected artistic styles, resulting in 80 hand-painted references (four gloss levels $\times$ five styles $\times$ four artists). Figure 3 (rows 2 to 6) shows the paintings made by one of the artists; please refer to the supplementary material for additional details.

Since our work involves the application of these styles to different geometries (Section 5) and the creation of a larger dataset (Section 6, Training and Evaluation Datasets), producing such a vast number of hand-painted drawings would be prohibitively time-consuming. To overcome this, we utilized StyLit [FJL*16] (available at `https://github.com/jamriska/ebsynth`), a method that allows us to transfer the artistic styles from the reference paintings of a render to target renders of different geometries, being robust even under slight changes in illumination (i.e., light position) between the reference painting and the target render. StyLit transfers patches from the original hand-painted references to the target objects, ensuring that the unique features of each style and artist are preserved. StyLit takes as input Light Path Expressions (LPEs) of both the reference and target renders: direct illumination maps to emphasize the contrast between lighted areas and shadows; specular maps to provide a proper stylization of highlights; and maps with the first and second diffuse bounces to emphasize details in shadows. In addition, StyLit also needs a binary mask to highlight the object's outline and an edge map (computed using the Canny detector) to preserve the semantics of the target geometry in the stylized render. We generated the LPEs (for both the reference and target render) and the binary mask using Mitsuba 3. By leveraging this technique, we were able to successfully transfer the original artistic styles to different geometries, which we use later in our study and for the creation of our dataset.

5. Gloss Perception in Painterly Depictions of Materials

To automatically annotate our dataset with text descriptions of gloss, we need to understand how this perceptual attribute is perceived in all five styles (i.e., oil painting, watercolor, ink pen, charcoal, and soft crayon) present in our dataset. Therefore, in this section, our goal is to determine whether gloss perception in the painterly depictions of our dataset remains consistent with that in their photorealistic counterparts.

Previous studies have investigated material perception across different visual media. For instance, Zuijlen et al. [vZPW20] analyzed how humans perceive high-level perceptual attributes, such as gloss, roughness, or hardness in paintings, and suggested that material perception operates independently of the medium of representation (i.e., paintings and photos). The work of Delanoy et al. [DSMG21] yielded similar conclusions, finding that our perception of materials in painterly depictions is linked to similar visual cues that we use to compute the perceived gloss in photorealistic renders. Building on these findings, we conducted a study to investigate how gloss perception is affected by different hand-drawn artistic styles (i.e., oil painting, watercolor, ink pen, charcoal, and soft crayon) in comparison to their photorealistic counterparts.

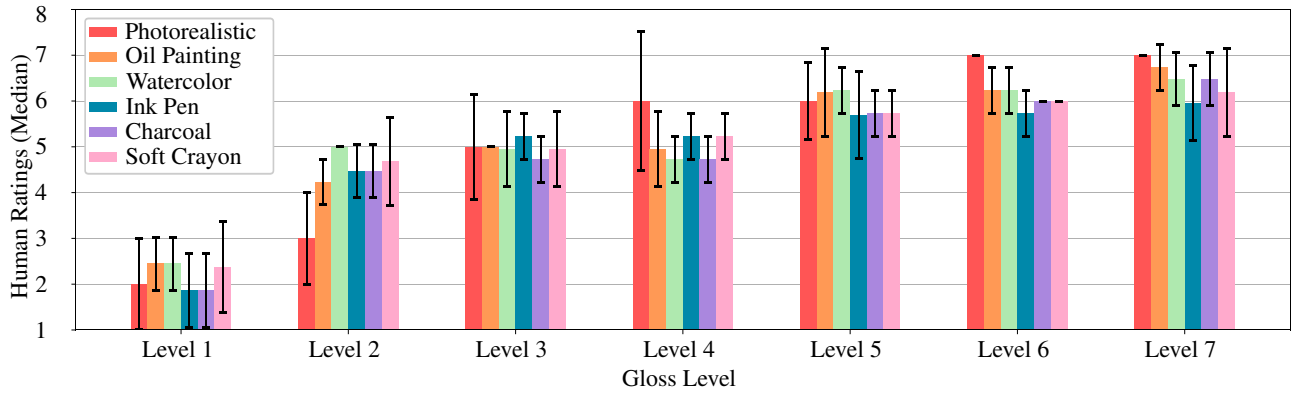


Figure 5: *Results of the seven-point Likert scale ratings for gloss, as assigned to the photorealistic renders and their stylized counterparts, averaged across all participants. Gloss levels are shown from left to right, with each artistic style represented by a different color. Error bars indicate the standard deviation, while bars without error reflect complete agreement among users.*

Stimuli For analyzing how the hand-drawn artistic styles present in our dataset affect human perception of gloss, we selected the *blob* geometry following the work of Vangorp et al. [VLD07], which suggests that this geometry is more suitable for judging the appearance of materials. We generate seven renders of the *blob* geometry under an area light placed on the top right corner of the scene by varying the roughness parameter r of the *Disney's Principled BSDF* [BS12, Bur15] within the set of values $\{0.0, 0.12, 0.24, 0.35, 0.47, 0.58, 0.7\}$, with a fixed albedo of 0.18. We styled each of the seven renders using StyLit [FJL*16], taking as reference the spheres painted by each of the four artists (Section 4). The photorealistic renders were stylized using the sphere paintings (see Figure 3, rows two through six) according to the following mapping:

- For the roughness value $r = 0.7$, we use the sphere paintings with the same roughness value.
- For the roughness value $r \in \{0.47, 0.58\}$, we use the sphere paintings with the roughness value $r = 0.47$.
- For a roughness value $r \in \{0.12, 0.24, 0.35\}$, we use the sphere paintings with the roughness value $r = 0.24$.
- Last, for a roughness value $r = 0.0$, we use the sphere paintings with the same roughness value.

We also apply histogram matching from the photorealistic renders to their stylized versions so that both representation media have the same tone of gray and do not influence gloss perception. In Figure 4 we show a render of the blob geometry (with a roughness value $r = 0.12$) and its stylized versions generated from the paintings of one of the artists. In total, we collected 140 painterly depictions (seven gloss levels $\times$ five styles $\times$ four artists) plus the seven photorealistic renders (147 images in total) for the *blob* geometry.

Participants Five participants (1 female and 4 males, 24 to 29 years old) with experience in computer graphics took part in the study. All participants had normal or corrected-to-normal vision and were naive to the goal of the study.

Procedure Participants were asked to rate the gloss levels of both the photorealistic renders and their stylized versions on a seven-point Likert scale. Prior to the annotation process with the blob, each participant rated the gloss level of the spheres painted by each of the artists as a calibration. Then, each user annotated the images with the *blob* geometry under constant controlled viewing conditions (Standard-dynamic-range display and fixed lighting). Each image was manually annotated for gloss by five different subjects. We provide additional details in the supplementary material.

Results The participants' annotations resulted in a high agreement (Krippendorff's alpha [Kri11]= 0.72; 1.0 would indicate perfect agreement). Figure 5 shows the results of the collected human ratings in our study. We can see how the perception of gloss follows a similar trend for both photorealistic and painterly depictions. We hypothesize that the users focus on similar visual cues that are used to compute the perceived gloss in photorealistic renders, in agreement with the hypothesis of Delanoy et al. [DSMG21] and Zuijlen et al. [vZPW20]. Furthermore, we performed a statistical analysis to examine potential effects of style on perceived gloss. Given our repeated measures design, where the same users rated multiple images, we computed a mixed-effects model, treating users as a random effect to account for within-subject variability. Our analysis, with photorealistic as the reference style, did not detect a statistically significant effect of style on ratings (all $p > 0.05$). To further investigate potential differences, we conducted pairwise comparisons using Tukey's HSD test with a Bonferroni correction: the effect sizes (Cohen's d) were uniformly small, indicating minimal practical differences between styles.

6. Datasets and Training

Training ControlNet [ZRA23] requires large datasets composed of millions of images paired with text descriptions and condition images (e.g., edge maps). We first describe the dataset of painterly depictions, paired with weakly annotated text descriptions of their appearance and edge maps, that we used to train ControlNet, and describe our evaluation dataset used to assess our framework (Section 6.1). Finally, we provide the technical details of the training process (Section 6.2).

6.1. Training and Evaluation Datasets

Collecting large datasets of painterly depictions of varying appearance paired with text descriptions, directly painted by artists and annotated through user studies, is impractical. First, having artists manually paint all possible variations of geometries and illuminations would be extremely time-consuming. Second, annotating such a large number of images through user studies is not feasible.

To address these challenges, we leverage StyLit [FJL*16] to transfer artistic styles from the hand-painted reference spheres to more complex geometries under varying illuminations and gloss levels. Since our user study (Section 5) suggests that gloss perception remains consistent across our photorealistic and stylized representations, we propagate the gloss annotations obtained from the photorealistic renders to their corresponding painterly versions. This allows us to generate an extensive dataset for training, composed of 1,336,272 painterly depictions using weak labels of perceived gloss [GVSS*24] that we later use to derive text descriptions of their appearance.

To generate our data, we leverage StyLit [FJL*16] to transfer the artistic styles from our reference spheres (Section 4) into different geometries and illuminations. We include 41 different geometries of varying complexities and details, each viewed from three different points of view, and under four illuminations by rotating an area light around the object in different positions. As materials, we generate seven gloss levels using the *Disney's Principled BSDF* [BS12, Bur15], by varying the value of the roughness parameter r in the interval $[0.0, 0.7]$. This results in a total of 17,220 painterly depictions for each single artist (41 geometries $\times$ 3 points of view $\times$ 4 illuminations $\times$ 7 gloss levels $\times$ 5 styles).

Up to this point, our dataset is only composed of gray painterly depictions, as painted by the artists. Additionally, we introduce variations in color as a post-processing step by generating color versions of our paintings of the sphere using the colorize function available in *GIMP* [Gim], except for the charcoal style, which remains only in gray. We generate 23 colors by sampling the HSLuv color space, using different probability distributions for each channel to achieve uniform perceptual sampling of the space. Considering all four artists and color variations, this results in a total of 1,336,272 painterly depictions. Finally, to automatically generate condition images for training our framework, we compute edge maps from the painterly depictions using the Canny edge detector. These edge maps serve as conditioning inputs to guide the network in preserving the desired shape of the depictions. For further details on the construction of the dataset, refer to the supplementary material.

Text Annotation Process Following the approach of Butt et al.[BWVCW24], we associate each stylized render in our dataset with a text description in the following template: *"A [GL] [G] hand painted with a [C] [S] on a white background."* In this template, $[GL]$ represents the gloss level, $[G]$ refers to the geometry (shape) of the object, $[C]$ refers to the color used during stylization, and $[S]$ corresponds to the hand-drawn artistic style of the image (i.e., oil painting, watercolor, ink pen, charcoal, and soft crayon).

The geometry, color, and style are obtained directly from the parameters used for rendering each image (see supplementary material). However, determining the gloss level is more complex since the final perceived appearance of an object, especially its gloss, depends not only on the material but also on the geometry and illumination [Fle14, SCW*21, LSGM21]. Perceived gloss, despite being grounded in photogeometric features such as the properties of reflected highlights [MKA12], cannot be well captured by objective measures. Consequently, using only the material's gloss, as determined by the BSDF of the rendered images, is insufficient to capture the overall perceived gloss. To address this, we aim to assign a gloss label to each rendered image that accounts for the interaction of material, geometry, and illumination. However, manually annotating gloss for every combination of these factors across the dataset is impractical. To overcome this challenge, we compute weak gloss labels for each of the photorealistic renders by computing the skewness of the image removing the background using a mask, as proposed by Guerrero-Viu et al. [GVSS*24]. However, directly calculating these weak gloss labels for the stylized images using automated computations would be unreliable due to visual features like brushstrokes, which may distort image statistics. Based on the findings from our user study (Section 5), which indicate that the hand-drawn artistic styles in our dataset do not strongly affect gloss perception, we apply the same gloss label to all styles, using the gloss label from the corresponding photorealistic render. Finally, the numerical weak gloss labels (ranging from one to seven) are mapped to text descriptions with the following categories: very matte, matte, somewhat matte, semi-glossy (neutral), somewhat glossy, glossy, and very glossy. Additional details are provided in the supplementary material.

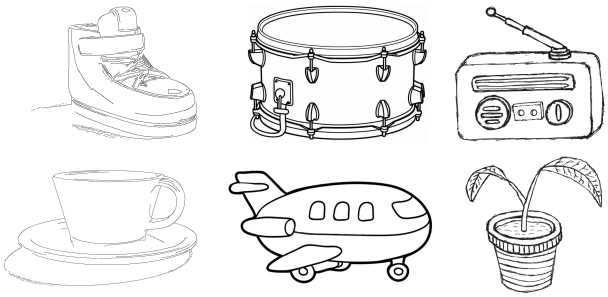


Figure 6: *Examples of condition images used in the evaluation dataset to assess the performance of our framework. The three types of condition inputs are shown from left to right: edge maps computed from photorealistic renders, clip arts, and hand-drawn sketches.*

Evaluation Dataset To evaluate our framework, we generated an additional set of painterly depictions never seen during training. We follow the same procedure as for generating the training dataset using: seven gloss levels, four illuminations by rotating an area light and 11 new geometries never seen during training. Therefore, we have a total of 308 photorealistic renders over which we apply the Canny edge detector to compute evaluation edge maps. For evaluation, we applied the Canny edge detector directly to these photorealistic images to generate the corresponding edge maps used as input conditions. This allows us to assess the framework's ability to convey key artistic features of each style when conditioned on

"A matte treasure chest in yellow oil painting" *"A matte sneaker in green watercolor"* *"A matte table lamp in red ink pen"* *"A glossy ligth bulb in gray charcoal"* *"A glossy tractor tire in blue soft crayon"*

Figure 7: *Painterly depictions computed from edge maps (top-left corners) never seen during training. Our framework generates lighting effects such as shadows or specular highlights following the semantics of the input edge map while depicting a visually compelling artistic appearance according to the input prompt.*

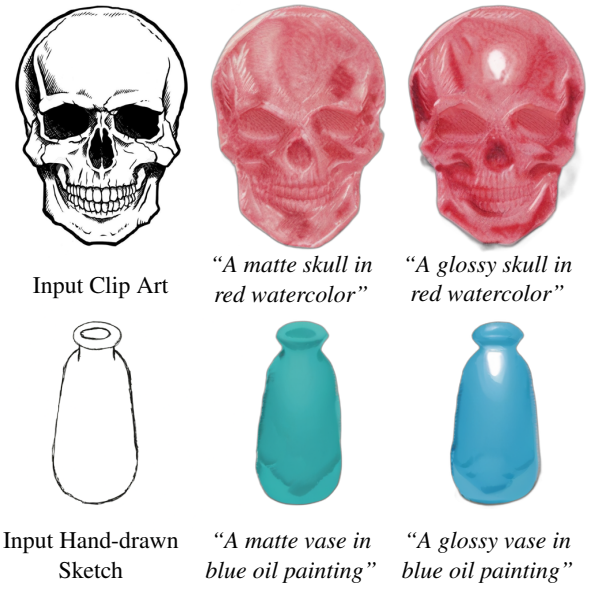
simplified inputs with reduced detail, unlike during training, where the edge maps were computed from the stylized images generated using StyLit [FJL*16]. For each render, we generate a stylized version using StyLit in order to compare it with the painterly depiction generated with our method (Section 7.2). We leverage the 23 colors (except for the charcoal style) and five styles (same four artists) present in the training dataset. Therefore, we have a total of 29,876 painterly depictions for a single artist. Considering all artists and color variations, this results in a total of 119,504 painterly depictions. We also collected a set of clip arts downloaded from the internet and hand-drawn sketches to evaluate the generalization capabilities of our framework. These are used directly as condition images (replacing the Canny edge maps) to test the network's ability to handle inputs with different visual characteristics and less precise structural guidance. Figure 6 shows a representative subset of the different input condition images present in our evaluation dataset. For further details, see supplementary material.

6.2. Training Details

To train our framework, we use the original ControlNet [ZRA23] loss function (unmodified). In terms of diffusion algorithms, given the input image $\mathbf{y}$, noise ϵ is progressively added to the image to produce a noisy image $\mathbf{y}_s$, where s represents the number of times noise is added. For the time step $\mathbf{s}$, the diffusion model ϵ_θ aims to predict the noise ϵ added to the image $\mathbf{y}_s$, given the condition image $\mathbf{x}$ and prompt $\mathbf{t}$. Thus the diffusion model minimizes the following loss function:

$$\mathcal{L} = \mathbb{E}_{\mathbf{y},\mathbf{s},\mathbf{x},\mathbf{t},\epsilon\sim\mathcal{N}(0,1)}\left[||\epsilon - \epsilon_\theta(\mathbf{y}_s,\mathbf{s},\mathbf{x},\mathbf{t})||_2^2\right], \quad (1)$$

where $\mathcal{N}$ denotes the normal distribution. We use as a backbone the initialization weights of a pre-trained version of Stable Diffusion XL (v1.0, available at https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0), without making any modifications to the original architecture. We train ControlNet using the ADAM optimizer [KB14] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate $\eta = 1 \times 10^{-5}$. We train our ControlNet with a batch size N = 8 (a total of 167,034 training steps, 1,336,272 images / 8 images per batch), an input resolution of 512×512 px., and a float-point precision of 16 bits. During the training process, we do not replace text prompts $\mathbf{t}$ with empty strings. All our experiments are computed using the PyTorch [PGM*19] library with cuDNN 12.2 executed in the Accelerate framework [GDW*22] on a NVIDIA RTX A6000 GPU. In total, training our framework takes around four days.



Input Clip Art *"A matte skull in red watercolor"* *"A glossy skull in red watercolor"*

Input Hand-drawn Sketch *"A matte vase in blue oil painting"* *"A glossy vase in blue oil painting"*

Figure 8: *Examples of painterly depictions synthesized from a clip art and a hand-drawn sketch (first column). Our framework can generate painterly depictions following the semantics of out-of-the-distribution condition images while depicting a matte appearance according to the input prompt (second column). Due to the absence of shadows, our framework assumes the illumination is frontal and generates consistent specular highlights when the gloss level is increased using the prompt (third column).*

7. Results

We first qualitatively evaluate the performance of our framework across several dimensions: (1) on edge maps computed from the

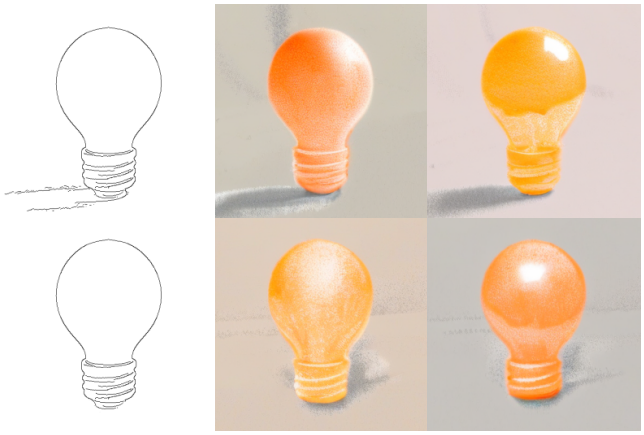


Figure 9: *First row: results when the shadow direction is present in the edge map. Our framework generates a matte or glossy appearance of the object according to the input prompt, while depicting shadows following the semantics of the input edge map. Second row: results in the absence of the shadow in the edge map. Our framework assumes that the illumination is frontal, depicting the shadow behind the object, while introducing specular highlights on the object's frontal surface when the gloss level is increased using the prompt (glossy).*

Figure 10: *Results with different configurations of the prompt that could be provided by the user: a prompt with a color not present in the training dataset (i.e., brown) and an insufficient prompt (i.e., without providing the style). Our framework can generate a visually compelling appearance despite the inconsistencies in the input prompt.*

photorealistic renderings (our evaluation dataset), (2) on out-of-distribution condition images such as clip arts and hand-drawn sketches, (3) in terms of consistency of the painterly depictions with respect to gloss level, (4) under different prompt configurations, (5) regarding consistency across the five hand-painted artistic styles included in our dataset, (6) under illumination changes, and (7) through an ablation study analyzing the impact of detail level in the condition image (Section 7.1). Next, we compare our results against StyLit [FJL*16] and other state-of-the-art diffusion models (Section 7.2). Finally, we conduct a user study to assess the fidelity of the generated painterly depictions with respect to the input text descriptions (Section 7.3).

7.1. Qualitative Evaluation of Our Framework

Consistency on Edge Maps We evaluate the performance of our framework on edge maps from our evaluation dataset (see Section 6.1). In Figure 7 we show how our framework infers the semantic information from edge maps of objects never seen during training, while depicting accurately lighting effects like shadows or specular highlights, and producing a visually compelling appearance according to the input prompts.

Generalization We have trained our framework with edge-map information; here we evaluate it on out-of-distribution clip arts and hand-drawn sketches instead. In Figure 8 we can see how our framework synthesizes the visual appearance of the input prompt, following the semantics of the objects represented in the input condition images. When requested in the prompt, our method introduces consistent specular highlights according to the object's shape as we can see in Figure 8 (second row), where the specular highlight elongates vertically on the vase's surface, and horizontally along its rim.

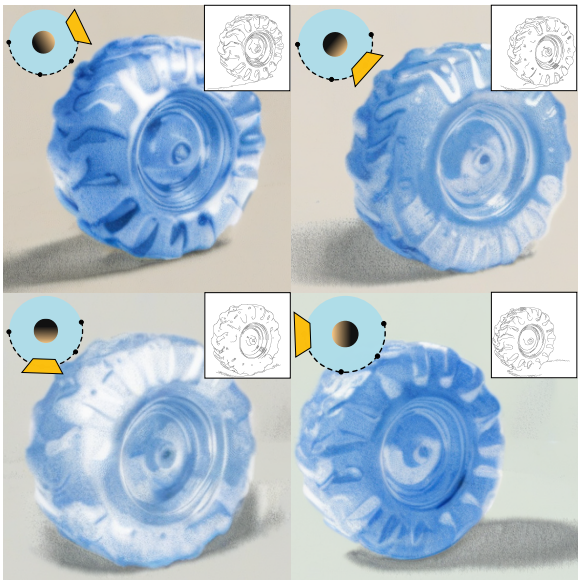
Consistency of Gloss Level We next evaluate the performance of our framework to generate a consistent artistic appearance according to the gloss level provided by the user (i.e., *matte* or *glossy*). Figure 9 shows painterly depictions of the same object when varying the desired gloss level through the input prompt. Our method introduces the specular highlights of the corresponding intensity according to the light direction suggested by the shadow in the input image (first row). In the absence of such shadow, our framework assumes that the object is illuminated by a frontal light (second row).

Prompt Configurations We evaluate our method with different configurations of the prompt that do not follow the format used during training, but could be provided by the user: a prompt with a color not present in the training dataset (for further details, see supplementary material), and an insufficient prompt that does not fully cover the desired appearance of the painterly depiction (e.g., no style is provided). Figure 10 shows how our framework performs reasonably well, synthesizing visually compelling painterly depictions.

Style Variations We assess the performance of our framework across the different styles in our dataset (i.e., oil painting, watercolor, ink pen, charcoal, and soft crayon). Figure 11 shows painterly depictions of the same object in all five styles and different colors. Our framework captures the key artistic traits of each style: brushstrokes in oil painting and watercolor; strokes in ink pen and stippled texture in charcoal and soft crayon.

Consistency under Illumination Changes We evaluate the performance of our method under the four illumination conditions present in our evaluation dataset, where an area light is rotated around the object while keeping the geometry fixed. As shown in Figure 12, our framework produces consistent painterly depictions that capture shadows, specular highlights, and soft shading effects, in accordance with the visual cues (such as shadow direction) present in the input condition image.

Figure 11: *Results of painterly depictions from a fixed clip art varying the style (from left to right: oil painting, watercolor, ink pen, charcoal and soft crayon) and color. We can observe how our framework synthesizes consistent painterly depictions for all styles in our dataset, according to the input prompt.*



"A glossy tractor tire in blue soft crayon"

Figure 12: *Results under the four illuminations in our evaluation dataset, created by rotating an area light around a fixed geometry. Illumination directions are shown in the top-left corners, and the corresponding edge maps are shown in the top-right corners; all results were generated using a fixed prompt. Our framework depicts shadows, specular highlights, and soft shading, consistent with the direction of the incident light.*

Detail Level in the Condition Image We conduct an ablation study by varying the detail level in the condition image. As shown in Figure 13, the model is able to produce plausible specular highlights even from a basic outline. When detail is sparse the placement of highlights often defaults to frontal illumination patterns (as discussed in Section 7.1, Consistency of Gloss Level). As more structure is added to the edge map, such as contours around specular regions or shadow boundaries, the output more closely matches the lighting and appearance of the original render. This suggests that while the model generalizes reasonably well with minimal input, it benefits significantly from richer structural information.

7.2. Comparison to Previous Methods

StyLit First, we compare our results with StyLiyt [FJL*16], which relies on extensive input including LPEs (obtained from rendered images of the 3D object), a binary mask, an edge map, and a sphere painted in the same style as the desired. In contrast, our method allows to synthesize painterly depictions from just an input condition image: an edge map, a clip art or a hand-drawn sketch (where no 3D information exists), and a text prompt. Thus our method does not require a stylized version of the reference render to guide the editing process, since this information is encoded in the input prompt. In Figure 14, we show a qualitative comparison of our method with StyLit, both using the full input set (StyLit Full), and reducing it to an input more similar to ours (StyLit Base). As we can see, the stylized results generated by StyLit strongly depend on the additional (3D-based) input information; given a suboptimal input more close to ours the quality of its results drops significantly (see the supplementary material for additional comparisons).

Other Diffusion Models We also compare our method against other state-of-the-art diffusion models. Specifically, we show results of ControlNet (v1.1) [ZRA23], T2I-Adapter [MWX*24], ControlNet++ [LYK*24], and a pre-trained version of Stable Diffusion XL conditioned by ControlNet but not trained on our dataset (ControlNet SDXL). Figure 15 shows the results using both a clip art and a hand-drawn sketch as input. We can see how existing methods fail to reproduce the outcome as described in the input prompt (for further comparisons see the supplementary material).

In addition, we also include quantitative comparisons. For this purpose, we compute the FID (Frechet Inception Distance) [HRU*17] and KID (Kernel Inception Distance) [PZZ22] to evaluate the quality of the generated painterly depictions, and compute the CLIP Score [HHF*21] to evaluate how well the generated painterly depictions matches the input text descriptions. For these evaluations, we randomly select 100 samples from our evaluation dataset, each consisting of a painterly depiction (generated using StyLit [FJL*16]), an associated edge map, and a text description. We use the edge maps and text prompts to generate painterly depictions with our method and with baseline diffusion models. The FID and KID scores are computed between the outputs of each model and the corresponding StyLit references. To calculate the CLIP Score, we measure the similarity between the generated images and their corresponding text prompts. Table 1 shows that our framework

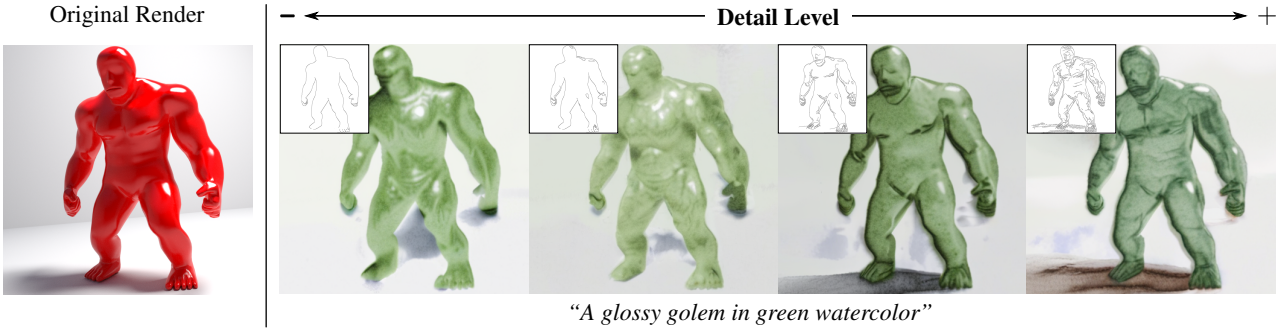


Figure 13: *Results for a fixed geometry (original render) and prompt, by increasing (from left to right) the detail level present in the condition image (top-left corners). Even with minimal input (such as a basic outline), our framework produces plausible specular highlights, although the lack of detail often results in generic frontal placement. Adding structural cues, such as contours around specular regions or shadows, improves alignment with the visual cues present in the original render.*

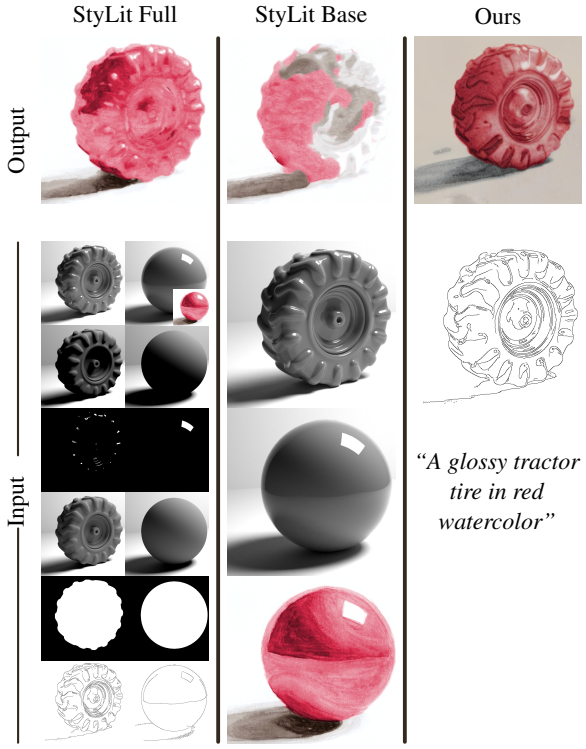


Figure 14: *Top row: comparison of painterly depictions generated using StyLit [FJL*16], both using the full set of needed inputs (StyLit Full) and reducing it to an input more similar to ours (StyLit Base); and our proposed method. StyLit Full input includes the target render, the reference render of a sphere, a sphere painted in the desired style, LPEs, binary masks and edge maps. StyLit Base input includes the target render, the reference render of a sphere and a sphere painted in the same style. In contrast, our framework's input only includes: an edge map and a prompt describing the desired output; yielding more accurate results.*

Table 1: *Quantitative comparison with state-of-the-art diffusion models, based on the mean scores over a subset of 100 painterly depictions from our evaluation dataset. Our framework more closely matches the distribution of the StyLit references (as indicated by lower FID and KID scores), while also achieving higher alignment with the input text descriptions (as measured by the CLIP Score).*

Method	FID ↓	KID ↓	CLIP Score ↓
ControlNet (v1.1)	176.07 ± 5.335	0.030 ± 0.000	0.300 ± 0.029
T2I-Adapter	181.37 ± 7.976	0.026 ± 0.000	0.320 ± 0.026
ControlNet++	182.02 ± 6.352	0.028 ± 0.000	0.310 ± 0.033
ControlNet SDXL	200.92 ± 6.473	0.041 ± 0.000	0.297 ± 0.033
Ours	**127.64 ± 6.327**	**0.009 ± 0.000**	**0.294 ± 0.032**

consistently outperforms the other diffusion models, supporting the conclusions drawn from our qualitative evaluation.

7.3. User Study

We conducted a user study to evaluate the quality of the generated painterly depictions in terms of their alignment with the input text descriptions.

Stimuli We generated 20 painterly depictions (10 matte and 10 glossy) using ControlNet (v1.1), T2I-Adapter, ControlNet++, ControlNet SDXL, and our framework, all conditioned on a variety of edge maps, clip arts, and hand-drawn sketches. Each question presented participants with a strip of five painterly depictions (one from each diffusion model) generated from the same condition image and text prompt. To help participants understand the intended artistic traits, we also showed a grayscale painting of a reference sphere created by one of the artists in the specified hand-drawn style (i.e., oil paint, watercolor, ink pen, charcoal, or soft crayon)

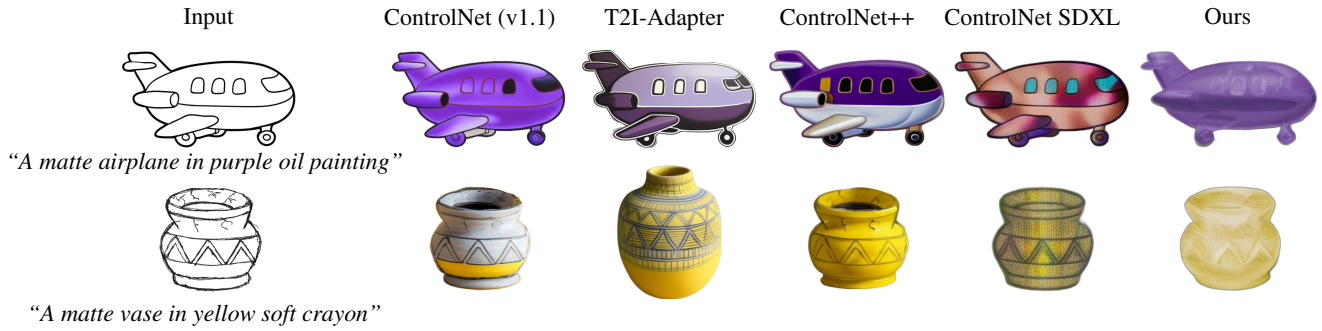


Figure 15: *Qualitative comparison to previous methods: ControlNet (v1.1) [ZRA23], T2I-Adapter [MWX*24], ControlNet++ [LYK*24], and pre-trained Stable Diffusion XL [PEL*23] conditioned with ControlNet but not trained on our dataset (ControlNet SDXL); and our method. We can observe how our method leads to much better painterly depictions, inferring the semantics from the input condition image, while depicting a visually compelling appearance (representing the strokes and aesthetic qualities) according to the input prompt.*

and gloss level (i.e., matte or glossy), as described in the prompt (see Section 4). We randomized the order of the strips and the left-to-right placement of the generated images within each strip. Participants were asked to rank the five painterly depictions from most to least faithful to the text description (first to fifth). This ranking captures both text alignment and perceived visual quality, since images that appear overly photorealistic may not be perceived as belonging to a hand-drawn artistic style. Additional details about the experimental setup can be found in the supplementary material.

Participants A total of 27 participants (10 females and 17 males, 22 to 60 years old) took part in the study. All participants had normal or corrected-to-normal vision.

Results Overall, our method was preferred over all other diffusion models, as measured by the mean percentage of pairwise wins, i.e., how often our method was ranked higher than each competing model across all comparisons (see Table 2, second column). We also report the rank product:

$$RP(g) = \left(\prod_{k=1}^{K} r_{g,k} \right)^{\frac{1}{K}}. \quad (2)$$

Where $r_{g,k}$ is the average ranking assigned to diffusion model g by participants for question k, with $K = 20$ questions in total. As shown in Table 2 (third column), our method achieves the lowest rank product score, indicating that it was consistently rated as the most faithful to the input descriptions (lower values denote better rankings).

8. Discussion and Future Work

Our work allows to synthesize painterly depictions of objects in different hand-drawn artistic styles. We have also introduced a large non-photorealistic dataset of painterly depictions automatically annotated with text descriptions of their appearance, along with their corresponding edge maps, which suffice to condition a framework based on Stable Diffusion XL.

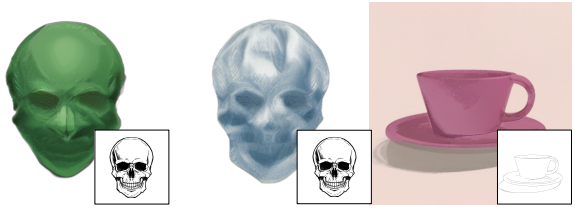


Figure 16: *Example of three failure cases of our framework. (a) No geometric information is provided in the input prompt; (b) the condition image and geometric information provided in the input prompt are not aligned; and (c) a style not present in our dataset.*

Table 2: *User study results comparing our framework against state-of-the-art diffusion models in terms of alignment with the input text descriptions. We report the win rate (percentage of pairwise comparisons where our method is ranked higher compared to the others) and the rank product (RP), which reflects overall user preference across all questions. Our painterly depictions are consistently preferred, with the highest win rate and the lowest RP score, indicating strong alignment with the intended descriptions.*

Method	Preference ↑	RP ↓
ControlNet (v1.1)	85.00 %	3.3293
T2I-Adapter	79.44 %	2.7807
ControlNet++	85.74 %	3.3913
ControlNet SDXL	87.04 %	3.5083
Ours	-	**1.5651**

While other approaches rely on multiple sources of information (including Light Path Expressions plus a painted exemplar of the desired output [FJL*16]), our method takes a single condition image as input, including out-of-distribution input such as clip arts or hand-drawn sketches; and a text description. We have shown its performance across a variety of objects of varying complexity, dif-

Figure 17: *Example of controlling the position of specular highlights and shadows by explicitly drawing shadow outlines in the condition image (first row). In contrast, the second and third rows show failure cases where the position of specular highlights and shadows, respectively, is specified only in the prompt without corresponding cues in the condition image.*

ferent styles, reflectance levels, illuminations, and even suboptimal text prompts and condition images defining the desired outcome.

However, our method is not free of limitations. As shown in Figure 16 (a), our framework is trained using prompts that include the label of the depicted geometry in the condition image. When this information is omitted, the model's ability to accurately delineate object contours diminishes. Performance also degrades when there is a mismatch between the geometry mentioned in the prompt and the condition image (Figure 16 (b)), leading to a loss of details. This issue could be mitigated by randomly replacing geometry labels with empty strings during training, encouraging the model to infer object identity from the condition image alone, as suggested in ControlNet [ZRA23]. On the other hand, our framework does not generalize well to more complex artistic styles beyond those included in our dataset, such as impressionism, cubism, constructivism, or surrealism (Figure 16 (c)). As a data-driven approach, the quality of our results is limited by the training data. For instance, painterly depictions in the training dataset exhibit artifacts on flat or concave surfaces, since StyLit [FJL*16] struggles in these cases due to the use of a sphere as exemplar. These artifacts tend to be encoded into the diffusion model resulting in a loss of quality, further accentuated by the known difficulty of CNN-based models to preserve low-level details. This limitation could be mitigated by augmenting the training data and incorporating broader style datasets (i.e., [SLK*19]) or hand-drawn examples collected directly from artists. Moreover, when working with edge maps, the Canny detector lacks robustness on low-contrast images, leading to erroneous stylization, which could be mitigated by using more advanced line detection methods [VGJMR08, FH23, PBL*23]. On the other hand, results on clip arts or hand-drawn sketches may exhibit underlying shadows, which could be minimized by using more precise background masking techniques. Manually drawing shadow outlines in the condition image enables control over shadow and specular highlight positions (Figure 17, first row). However, the diffusion model is not able to reliably adjust these visual cues using the text prompt alone (Figure 17, second and third rows), likely due to bias introduced by the training data and specifically, the lack of explicit descriptions of shadow direction or highlight placement in the training prompts. To address this limitation, training prompts could be augmented with controlled text descriptions that explicitly encode the direction of shadows and the position of specular highlights.

We hope our work inspires promising directions for future research. For simplicity, we have focused on two of the most prominent high-level perceptual attributes: gloss and color, as they are widely studied and easily understood by novice artists. Nevertheless, other perceptual traits could also be explored. Our pipeline is designed to be easily extendable, allowing the inclusion of new styles with minimal manual effort. For instance, to support other styles or multicolored objects, only a small set of spheres painted in such styles is required. These can then be propagated using StyLit to generate a large training dataset, after which the model can be retrained following the same procedure described in this work. We hope this modular design will help foster promising directions for future work. Furthermore, our method could be extended to stylize video sequences by using our diffusion model as a backbone, while incorporating mechanisms to enforce temporal consistency [GBTBD, MHV*23]. Due to memory limitations, we trained our framework at a resolution of 512×512 px., which pushed the limits of our available resources. However, we believe the method should scale to higher resolutions, provided sufficient computational resources are available.

Finally, new interaction modalities beyond text prompts could be integrated into our framework for a better user experience. One such example is the recently proposed TexSliders system by Guerrero-Viu et al. [GVHR*24], which offers intuitive, slider-based control over texture attributes.

Acknowledgments

This work was supported by the project PID2022-141539NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU; by the Government of Aragon's Departamento de Ciencia, Universidad y Sociedad del Conocimiento through the Reference Research Group "Graphics and Imaging Lab" (ref. T34_23R); and by the Government of Aragon's Departamento de Educación, Ciencia y Universidades through the project "HUMAN-VR: Development of a Computational Model for Virtual Reality Perception" (PROY_T25_24). J. Daniel Subias was supported by the CUS/702/2022 predoctoral grant. We thank Sergio Cartiel for his help with the figures; the members of the Graphics and Imaging Lab for their interesting discussions and help with proofreading; and all the participants of the user studies. We also thank Ignacio Moreno, Edurne Bernal, Erika Dolz, Noa Fernandez and Lucia Barriendo for their paintings, used to create our dataset.

References

[AKLM18] ADAMS W. J., KUCUKOGLU G., LANDY M. S., MANTIUK R. K.: Naturally glossy: Gloss perception, illumination statistics, and tone mapping. *Journal of Vision 18*, 13 (2018), 4–4.

[And11] ANDERSON B. L.: Visual perception of materials and surfaces. *Current biology 21*, 24 (2011), R978–R983.

[BBSM*25] BERNAL-BERDUN E., SERRANO A., MASIA B., GADELHA M., HOLD-GEOFFROY Y., SUN X., GUTIERREZ D.: Precisecam: Precise camera control for text-to-image generation, 2025. URL: https://arxiv.org/abs/2501.12910, arXiv:2501.12910.

[BKTS06] BOUSSEAU A., KAPLAN M., THOLLOT J., SILLION F. X.: Interactive watercolor rendering with temporal coherence and abstraction. In *Proceedings of the 4th international symposium on Non-photorealistic animation and rendering* (2006), pp. 141–149.

[BLV*10] BÉNARD P., LAGAE A., VANGORP P., LEFEBVRE S., DRETTAKIS G., THOLLOT J.: A dynamic noise primitive for coherent stylization. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 1497–1506.

[BOD*13] BOUSSEAU A., O'SHEA J. P., DURAND F., RAMAMOORTHI R., AGRAWALA M.: Gloss perception in painterly and cartoon rendering. *ACM Transactions on Graphics (TOG) 32*, 2 (2013), 1–13.

[BRL97] BHUSHAN N., RAO A. R., LOHSE G. L.: The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images. *Cognitive Science 21*, 2 (1997), 219–246.

[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).

[BWVCW24] BUTT M. A., WANG K., VAZQUEZ-CORRAL J., WEIJER J. V. D.: Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement, 2024.

[CJP*23] CHEN B., JINDAL A., PIOVARČI M., WANG C., SEIDEL H.-P., DIDYK P., MYSZKOWSKI K., SERRANO A., MANTIUK R. K.: The effect of display capabilities on the gloss consistency between real and virtual objects. In *SIGGRAPH Asia 2023 Conference Papers* (2023), pp. 1–11.

[CK15] CHADWICK A., KENTRIDGE R.: The perception of gloss: A review. *Vision Research 109* (2015), 221–235.

[CLY*17] CHEN D., LIAO J., YUAN L., YU N., HUA G.: Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1105–1114.

[CMK*14] CIMPOI M., MAJI S., KOKKINOS I., MOHAMED S., VEDALDI A.: Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 3606–3613.

[CPBD23] CONDOR J., PIOVARCI M., BICKEL B., DIDYK P.: Gloss-aware color correction for 3d printing. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023), pp. 1–11.

[CS16] CHEN T. Q., SCHMIDT M.: Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337* (2016).

[DGVG*23] DESCHAINTRE V., GUERRERO-VIU J., GUTIERREZ D., BOUBEKEUR T., MASIA B.: The visual language of fabrics. *ACM Trans. Graph. 42*, 4 (jul 2023).

[DLGM22] DELANOY J., LAGUNAS M., GUTIERREZ D., MASIA B.: A generative framework for image-based editing of material appearance using perceptual attributes. *Computer Graphics Forum 41*, 1 (2022), 453–464.

[DS02] DECARLO D., SANTELLA A.: Stylization and abstraction of photographs. *ACM transactions on graphics (TOG) 21*, 3 (2002), 769–776.

[DSMG21] DELANOY J., SERRANO A., MASIA B., GUTIERREZ D.: Perception of material appearance: a comparison between painted and rendered images. *Journal of Vision 21*, 5 (May 2021), 16–16.

[DTD*21] DENG Y., TANG F., DONG W., HUANG H., MA C., XU C.: Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), vol. 35, pp. 1210–1217.

[FCC*19] FUTSCHIK D., CHAI M., CAO C., MA C., STOLIAR A., KOROLEV S., TULYAKOV S., KUCERA M., SỲKORA D.: Real-time patch-based stylization of portraits using generative adversarial network. *Expressive 2019* (2019), 33–42.

[FEHF09] FARHADI A., ENDRES I., HOIEM D., FORSYTH D.: Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 1778–1785.

[FH23] FANG C.-Y., HAN X.-F.: Joint geometric-semantic driven character line drawing generation. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval* (2023), pp. 226–233.

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

[FJS*17] FIŠER J., JAMRIŠKA O., SIMONS D., SHECHTMAN E., LU J., ASENTE P., LUKÁČ M., SỲKORA D.: Example-based synthesis of stylized facial animations. *ACM Transactions on Graphics (TOG) 36*, 4 (2017), 1–11.

[Fle14] FLEMING R. W.: Visual perception of materials and their properties. *Vision research 94* (2014), 62–75.

[Fle17] FLEMING R. W.: Material perception. *Annual review of vision science 3* (2017), 365–388.

[GBTBD] GEYER M., BAR-TAL O., BAGON S., DEKEL T.: Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*.

[GDW*22] GUGGER S., DEBUT L., WOLF T., SCHMID P., MUELLER Z., MANGRULKAR S., SUN M., BOSSAN B.: Accelerate: Training and inference at scale made simple, efficient and adaptable., 2022.

[GEB15] GATYS L. A., ECKER A. S., BETHGE M.: A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

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

[GOS*10] GED G., OBEIN G., SILVESTRI Z., LE ROHELLEC J., VIÉNOT F.: Recognizing real materials from their glossy appearance. *Journal of vision 10*, 9 (2010), 18–18.

[GVHR*24] GUERRERO-VIU J., HASAN M., ROULLIER A., HARIKUMAR M., HU Y., GUERRERO P., GUTIERREZ D., MASIA B., DESCHAINTRE V.: Texsliders: Diffusion-based texture editing in clip space. In *ACM SIGGRAPH 2024 Conference Papers* (2024).

[GVSS*24] GUERRERO-VIU J., SUBIAS J. D., SERRANO A., STORRS K. R., FLEMING R. W., MASIA B., GUTIERREZ D.: Predicting perceived gloss: Do weak labels suffice? *Computer Graphics Forum 43*, 2 (2024), e15037.

[Hae90] HAEBERLI P.: Paint by numbers: Abstract image representations. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques* (1990), pp. 207–214.

[HE04] HAYS J., ESSA I.: Image and video based painterly animation. In *Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering* (2004), pp. 113–120.

[Her98] HERTZMANN A.: Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques* (1998), pp. 453–460.

[HHF*21] HESSEL J., HOLTZMAN A., FORBES M., BRAS R. L., CHOI Y.: Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).

[HJO*01] HERTZMANN A., JACOBS C., OLIVER N., CURLESS B., SALESIN D.: Image analogies. *Proceedings of ACM SIGGRAPH 2001* (06 2001).

[HRU*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems 30* (2017).

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14* (2016), Springer, pp. 694–711.

[JSR*22] JAKOB W., SPEIERER S., ROUSSEL N., NIMIER-DAVID M., VICINI D., ZELTNER T., NICOLET B., CRESPO M., LEROY V., ZHANG Z.: Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org.

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[KK87] KAMADA T., KAWAI S.: An enhanced treatment of hidden lines. *ACM Transactions on Graphics (TOG) 6*, 4 (1987), 308–323.

[KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4401–4410.

[KPNM19] KUMAR M. P., POORNIMA B., NAGENDRASWAMY H., MANJUNATH C.: A comprehensive survey on non-photorealistic rendering and benchmark developments for image abstraction and stylization. *Iran Journal of Computer Science 2* (2019), 131–165.

[Kri11] KRIPPENDORFF K.: Computing krippendorff's alpha-reliability. URL: https://api.semanticscholar.org/CorpusID:59901023.

[LDX*19] LIU M., DING Y., XIA M., LIU X., DING E., ZUO W., WEN S.: Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 3673–3682.

[Lit97] LITWINOWICZ P.: Processing images and video for an impressionist effect. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (1997), pp. 407–414.

[LSGM21] LAGUNAS M., SERRANO A., GUTIERREZ D., MASIA B.: The joint role of geometry and illumination on material recognition. *Journal of Vision 21*, 2 (2021), 2–2.

[LXJ12] LU C., XU L., JIA J.: Combining sketch and tone for pencil drawing production. In *Proceedings of the symposium on non-photorealistic animation and rendering* (2012), pp. 65–73.

[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).

[Mei96] MEIER B. J.: Painterly rendering for animation. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques* (1996), pp. 477–484.

[MHV*23] MOLAD E., HORWITZ E., VALEVSKI D., ACHA A. R., MATIAS Y., PRITCH Y., LEVIATHAN Y., HOSHEN Y.: Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329* (2023).

[MKA12] MARLOW P. J., KIM J., ANDERSON B. L.: The perception and misperception of specular surface reflectance. *Current Biology 22*, 20 (2012), 1909–1913.

[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.

[PBL*23] PAUTRAT R., BARATH D., LARSSON V., OSWALD M. R., POLLEFEYS M.: Deeplsd: Line segment detection and refinement with deep image gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 17327–17336.

[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.

[PGM*19] PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KOPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J., CHINTALA S.: Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (2019), Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., (Eds.), vol. 32, Curran Associates, Inc.

[PZZ22] PARMAR G., ZHANG R., ZHU J.-Y.: On aliased resizing and surprising subtleties in gan evaluation. In *CVPR* (2022).

[QZY*23] QIN C., ZHANG S., YU N., FENG Y., YANG X., ZHOU Y., WANG H., NIEBLES J. C., XIONG C., SAVARESE S., ET AL.: Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147* (2023).

[RBL*21] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10674–10685.

[RDN*22] RAMESH A., DHARIWAL P., NICHOL A., CHU C., CHEN M.: Hierarchical text-conditional image generation with clip latents. *arXiv abs/2204.06125* (2022).

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (2015), Springer, pp. 234–241.

[SCW*21] SERRANO A., CHEN B., WANG C., PIOVARČI M., SEIDEL H.-P., DIDYK P., MYSZKOWSKI K.: The effect of shape and illumination on material perception: model and applications. *ACM Transactions on Graphics 40*, 4 (2021), 1–16.

[SJT*19] SÝKORA D., JAMRIŠKA O., TEXLER O., FIŠER J., LUKÁČ M., LU J., SHECHTMAN E.: StyleBlit: Fast example-based stylization with local guidance. *Computer Graphics Forum 38*, 2 (2019), 83–91.

[SL23] SUBÍAS J. D., LAGUNAS M.: In-the-wild Material Appearance Editing using Perceptual Attributes. *Computer Graphics Forum* (2023).

[SLK*19] SHUGRINA M., LIANG Z., KAR A., LI J., SINGH A., SINGH K., FIDLER S.: Creative flow+ dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[SMGG01] SLOAN P.-P., MARTIN W., GOOCH A., GOOCH B.: The lit sphere: A model for capturing npr shading from art. *No description on Graphics interface 2001* (07 2001).

[SWHS97] SALISBURY M. P., WONG M. T., HUGHES J. F., SALESIN D. H.: Orientable textures for image-based pen-and-ink illustration. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (1997), pp. 401–406.

[VGJMR08] VON GIOI R. G., JAKUBOWICZ J., MOREL J.-M., RANDALL G.: Lsd: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence 32*, 4 (2008), 722–732.

[VLD07] VANGORP P., LAURIJSSEN J., DUTRÉ PH.: The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2007) 26*, 3 (July 2007), 77:1–77:9.

[vLv*23] ŠUBRTOVÁ A., LUKÁČ M., ČECH J., FUTSCHIK D., SHECHTMAN E., SÝKORA D.: Diffusion image analogies. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023), p. 79.

[vZPW20] VAN ZUIJLEN M. J. P., PONT S. C., WIJNTJES M. W. A.: Painterly depiction of material properties. *Journal of Vision 20*, 7 (07 2020), 7–7.

[WFGS07] WINNEMOLLER H., FENG D., GOOCH B., SUZUKI S.: Using npr to evaluate perceptual shape cues in dynamic environments. In *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering* (2007), pp. 85–92.

[WKO12] WINNEMÖLLER H., KYPRIANIDIS J. E., OLSEN S. C.: Xdog: An extended difference-of-gaussians compendium including advanced image stylization. *Computers & Graphics 36*, 6 (2012), 740–753.

[WTM20] WU C., TIMM M., MAJI S.: Describing textures using natural language. In *European Conference on Computer Vision* (2020), Springer, pp. 52–70.

[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.

[ZZ11] ZHAO M., ZHU S.-C.: Portrait painting using active templates. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering* (2011), pp. 117–124.

[ZZXZ09] ZENG K., ZHAO M., XIONG C., ZHU S. C.: From image parsing to painterly rendering. *ACM Trans. Graph. 29*, 1 (2009), 2–1.