Preserve Anything: Controllable Image Synthesis with Object Preservation

Prasen Kumar Sharma Neeraj Matiyali Siddharth Srivastava Gaurav Sharma
Typeface
{prasen.sharma, neeraj.matiyali, siddharth, gaurav}@typeface.ai

1. Preliminaries

Text-to-image (T2I) synthesis methods, such as Stable Diffusion [13], generate images from textual descriptions (c^t) using diffusion models. These models operate in either the pixel or latent domain, with latent-space methods being computationally more efficient. Our work leverages Stable Diffusion, which employs a latent-space formulation for scalable and effective image synthesis. Below, we briefly outline the key components of Stable Diffusion and ControlNet [19], a framework enabling controlled image generation through additional task-specific inputs.

In diffusion models, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is incrementally added to an initial image x_0 to produce a noisy sample x_t at timestep t, as defined by:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad (1)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, following a variance schedule $\{\beta_t\}$. A denoising neural network ϵ_θ is trained to predict the noise ϵ by minimizing the objective:

$$\mathcal{L} = \mathbb{E}_{x_0, t, \epsilon \sim \mathcal{N}(0,I)} \left[\left\| \epsilon - \epsilon_\theta(x_t, t, c^t) \right\|_2^2 \right]. \quad (2)$$

Once the denoising network is trained, starting from some random noise $x_T \sim \mathcal{N}(0, I)$, it can be used to sample an image x_0 from the learned distribution by iteratively refining x_t:

$$x_{t-1} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t, c^t) \right) + \sigma_t \epsilon', \quad (3)$$

where $\epsilon' \sim \mathcal{N}(0, I)$ and $\sigma_t^2 = \beta_t$. Intermediate estimations of x_0 can also be obtained at any timestep t using:

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t, c^t)}{\sqrt{\bar{\alpha}_t}}. \quad (4)$$

In latent diffusion models (LDMs) (*e.g.*, Stable Diffusion), the forward and reverse diffusion process is done on latent features maps $z = \mathcal{E}_{\text{VAE}}(x)$ encoded by a pretrained autoencoder instead of RGB pixels x. The denoised latent representation z_0 is decoded into the final image using the decoder $x_0 = \mathcal{D}_{\text{VAE}}(z_0)$.

ControlNet [19] extends the controllability of large scale pretrained T2I LDMs process by introducing additional task-specific conditions (c^f), such as Canny edges, depth maps, or segmentation maps, which encode structural cues for the target image. The training objective of ControlNet incorporates these additional conditions:

$$\mathcal{L} = \mathbb{E}_{x_0, t, c^t, c^f, \epsilon \sim \mathcal{N}(0,I)} \left[\left\| \epsilon - \epsilon_\theta(x_t, t, c^t, c^f) \right\|_2^2 \right]. \quad (5)$$

This enables fine-grained control over the image generation process, allowing for precise alignment of the target object and structural layout in synthesized images.

2. Network Details

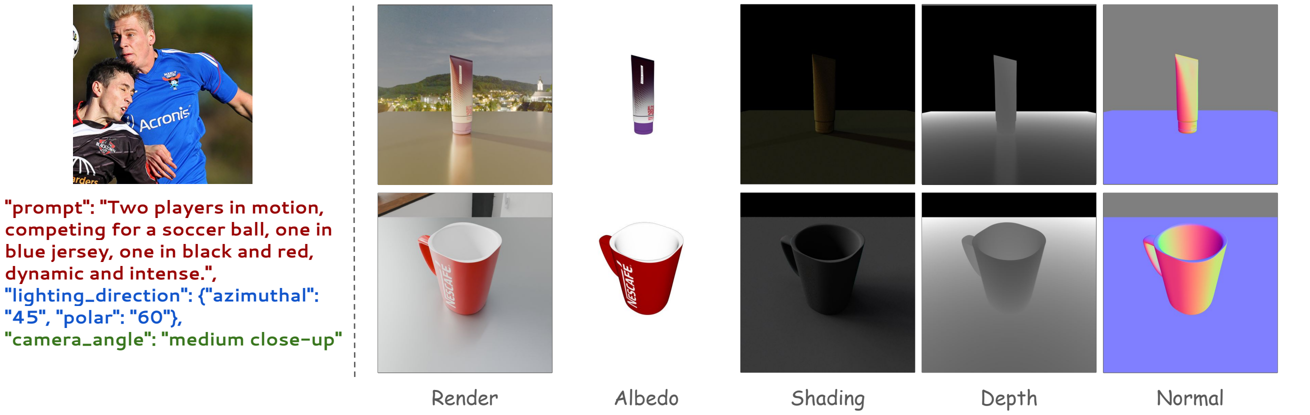
High-Frequency Overlay. The goal of this module is to re-establish high-frequency details in the target object of the synthesized image (J) from the source image (I). To achieve this, we employ a simple mechanism that decomposes an image into its low and high-frequency components. For any image P, The low-frequency component ($\ell_{lf}(P)$) is first extracted by applying a Gaussian blur with a kernel of size 17 × 17. The high-frequency component is then computed as:

$$\ell_{hf}(P) = P - \ell_{lf}(P). \quad (6)$$

We independently compute both the low and high-frequency components for the source and synthesized images, denoted as $\ell_{lf}(I)$, $\ell_{hf}(I)$, $\ell_{lf}(J)$, and $\ell_{hf}(J)$, respectively. Finally, we overlay the high-frequency components of the target object from the source image onto the synthesized image using its binary mask M as follows:

$$\hat{J} = M * \ell_{hf}(I) + (1 - M) * \ell_{hf}(J) + \ell_{lf}(J) \quad (7)$$

This approach enables high-frequency detail transfer with minimal computational overhead cf. to complex methods, such as Poisson blending [11], while maintaining a favorable trade-off between efficiency and visual quality.



"prompt": "Two players in motion, competing for a soccer ball, one in blue jersey, one in black and red, dynamic and intense.",
"lighting_direction": {"azimuthal": "45", "polar": "60"},
"camera_angle": "medium close-up"

Figure 1. **An illustration of our large-scale dataset.** (Left) shows an image from the real-world subset with its derived annotations (prompt and lighting cues for demonstration) from GPT4o. (Right) shows two samples from the curated 3D-rendered synthetic subset with their corresponding albedo, shading, depth, and surface normal maps.

3. Datasets

We introduce a novel large-scale dataset for controlled image synthesis (CIS), including real-world and 3D-rendered synthetic images, as outlined in Section 4.1 of the main manuscript. The following subsections provide a detailed description of the curated images and their corresponding annotations, for both real and synthetic image subsets.

3.1. Real-Image Subset

Real-world images are sourced from MS-COCO [7], Open-Imagesv7 [1], and FFHQ [6] datasets. However, instead of randomly selecting the images for training CIS models, we propose a filtering strategy to leverage images with better aesthetic quality by using LAION Aesthetic [14] score. Particularly, we utilize images with LAION Aesthetic scores above a threshold of 5.0. The threshold has been set based on the average LAION Aesthetic scores on MIT Adobe 5K [2] images. We also discard the grayscale images.

As reported in Section 4.1 of the main manuscript, for each selected image, we generate detailed annotations using GPT4o [10] as follows:

- five captions of varying lengths (20-50 words), describing salient objects and background elements,
- dominant light source in terms of spherical coordinates,
- camera orientation,
- objects in the scene with their detailed descriptions, and overall categorical counts,
- spatial relationships between different objects in the scene, *e.g.*, `The fence is behind the person`,
- action relationships, *e.g.*, `The person is holding the coffee cup`,
- technical details, such as light source with strength and direction, presence of shadows, color palette, depth of field, exposure, camera orientation in terms of roll, pitch, and yaw, and
- aesthetic scoring measuring overall composition, focus, clarity, memorability (which indicates "`how memorable an image is`"), timelessness, and emotions.

Our real-image subset consists of 240K images with the above-detailed annotations. These annotations are beneficial for solving diverse tasks. Quality prompts and lighting conditions are leveraged to enhance photorealism in CIS. Beyond image synthesis tasks, object descriptions with their spatial and action relationships could be used for efficient image understanding, and technical and aesthetic details for image quality assessment.

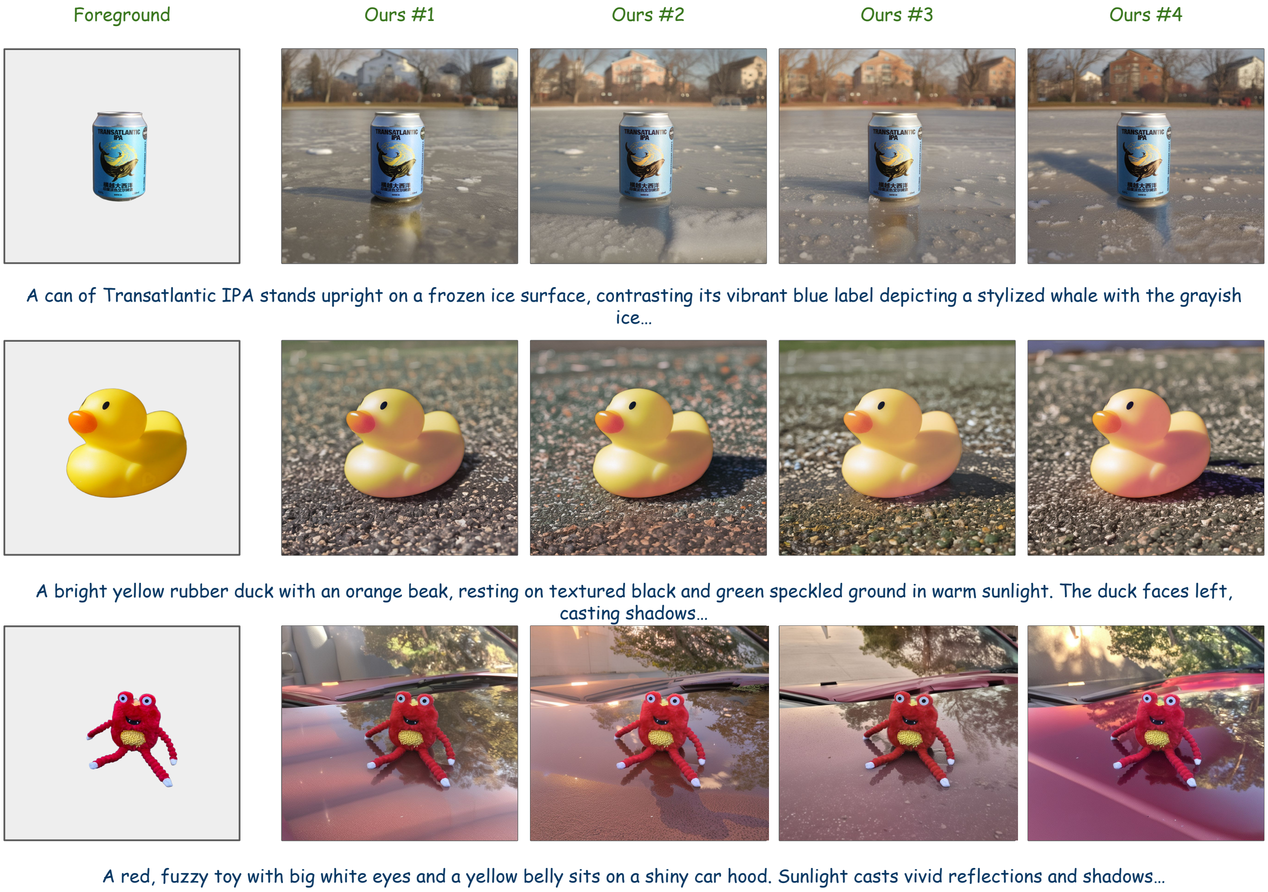
3.2. 3D Rendered Synthetic Subset

In addition to the real-image subset, we also construct a synthetic subset to control different natural aesthetics, such as lighting, and shadow, in the generated images. For this, we use Blender's [3] raytrace-based Cycles renderer to generate images of 3D assets. We collect different HDR environment textures from PolyHaven[1] for lighting the 3D assets while rendering. For each asset/environment pair, we vary camera orientation in terms of azimuthal angle from 0° to 360° and elevation angle from 45° to -15°.

For each render, we save maps capturing the essential 3D geometry, material, and illumination cues, which include albedo, diffuse shading, reflections, shadows, surface normals, and depth. These maps allow our method to learn lighting and shadow consistency and provide a basis for improving the realism of the generated images.

Figure 1 illustrates some samples from real-image and 3D-rendered subsets. (Left) shows an image from the real-world subset with its derived annotations (prompt and light-

[1] https://polyhaven.com/



A can of Transatlantic IPA stands upright on a frozen ice surface, contrasting its vibrant blue label depicting a stylized whale with the grayish ice...

A bright yellow rubber duck with an orange beak, resting on textured black and green speckled ground in warm sunlight. The duck faces left, casting shadows...

A red, fuzzy toy with big white eyes and a yellow belly sits on a shiny car hood. Sunlight casts vivid reflections and shadows...

Figure 2. **Consistency in natural aesthetics w.r.t. prompts.** Given a fixed background layout, observe how well the proposed method maintains the natural aesthetics such as lighting, casting shadows, and reflection w.r.t. input foreground and textual prompt, across various generations.

ing cues for demonstration) from GPT4o. The derived prompt describes the scene well, along with the accurate estimation of lighting coordinates and camera angle. (Right) shows two samples from the curated 3D-rendered synthetic subset with their corresponding albedo, shading, depth, and surface normal maps.

4. Quality Assessment

We employ the following image quality assessment metrics to measure the quality of generated images using Preserve Anything against existing works. These metrics evaluate image quality by assessing divergence from real images in the feature domain, prompt adherence, and visual appeal of the generated images. The metrics are briefly described as follows:

- The Fréchet Inception Distance (FID)[2] [5] metric assesses the quality of generated images by quantifying the divergence between feature distributions of real and generated images. It is commonly used to evaluate generative models, especially Generative Adversarial Networks (GANs) [4]. It is defined as the distance between two Gaussian distributions– one representing the real images and the other representing the generated images. These distributions are typically derived from a pre-trained Inception [15] model. A lower FID score indicates greater similarity between generated images and real images, reflecting the better performance of the generative model.
- The CLIP[3] [12, 18] scores assess the alignment between an image and a corresponding text prompt by measuring the similarity between their respective feature distributions. A higher CLIP score indicates better semantic correspondence between the image and the textual description.
- Neural Image Assessment (NIMA) [16] utilizes a con-

[2] https://github.com/GaParmar/clean-fid

[3] https://huggingface.co/docs/transformers/en/model_doc/clip

volutional neural network (CNN) to predict an image's aesthetic score on a scale of 1 to 10, which can be interpreted as a continuous value reflecting the image's quality or classified such as poor, average, excellent, *etc*.

- No-reference Quality Metric (NRQM) [8] is designed to predict the perceptual quality of images by analyzing various features, including sharpness, contrast, and other aesthetic or perceptual indicators. Specifically, it derives three types of low-level statistical features in both spatial and frequency domains, to measure the quantum of super-resolved artifacts, and learn a two-stage regression model to predict the quality of images, w/o referring to ground truth images.
- LAION Aesthetic[4] [14] (LAION-Aes) score estimates the aesthetic quality of an image using large-scale pre-trained models. It predicts the aesthetic appeal of an image, aligning with human perceptions of visual attractiveness, on a scale of 1 to 10, with 1 reflecting poor aesthetic quality. It is trained on a comprehensive set of images rated according to human judgment, such as Aesthetic Visual Analysis (AVA) [9]. The inline model is composed of simple linear layers on top of CLIP ViT/14 [12], and typically generates a score that reflects the overall visual appeal of an image, offering insight into how aesthetically pleasing it may be to human viewers.

We use pyiqa[5], a popular image quality assessment tool, for measuring NIMA and NRQM scores.

5. Ablation Study

We ablate the baseline "RGB Only-BLIP2" from Table 2 in paper by increasing the batch size from 16 to 128. With large batch size, the FID score (16.98) improves slightly (+0.33), while other scores show no significant gains. Larger batches may help stabilize gradients – evident from improved FID , but do not lead to better quality outputs.

6. More Results

The images generated using the proposed method demonstrate enhanced photorealism, preserving natural aesthetics such as lighting and shadows (see Figure 2), while the target objects are well-integrated with the background scene. Following recent work [17], and due to lack of established metrics to measure lighting and shadow consistency, we use GPT-4o as a visual critic to rate generated images on shadow strength, direction, and lighting adherence w.r.t. shadows. The average scores positively favor the pretraining. This shows (which is also evident from Figure 2) that pretraining with 3D rendered subset helps in generating images with consistent lighting and shadows.

[4] https : / / github . com / LAION - AI / aesthetic - predictor

[5] https://github.com/chaofengc/IQA-PyTorch

References

[1] Rodrigo Benenson and Vittorio Ferrari. From colouring-in to pointillism: revisiting semantic segmentation supervision. *CoRR*, abs/2210.14142, 2022. 2

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2

[3] Blender Online Community. Blender - a 3d modelling and rendering package, 2018. 2

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Shhamir Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 3

[5] 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. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 3

[6] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (12):4217–4228, 2021. 2

[7] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *CoRR*, abs/1405.0312, 2014. 2

[8] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-rolution. *Computer Vision and Image Understanding*, pages 1–16, 2017. 4

[9] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012. 4

[10] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2

[11] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, page 313–318, New York, NY, USA, 2003. Association for Computing Machinery. 1

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 4

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[14] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 4

[15] Christian Szegedy, Wei Liu, Yang Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 3

[16] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8): 3998–4011, 2018. 3

[17] Yu Tian, Yixuan Li, Baoliang Chen, Hanwei Zhu, Shiqi Wang, and Sam Kwong. Ai-generated image quality assessment in visual communication. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7392–7400, 2025. 4

[18] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 3

[19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 1