Learning Illumination Control in Diffusion Models

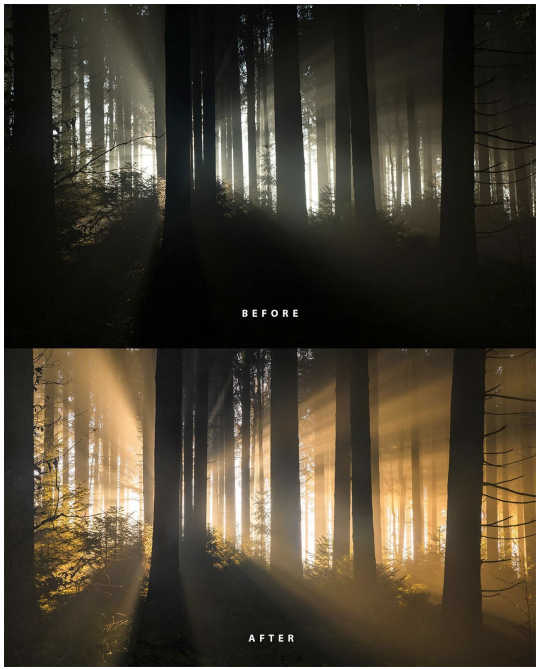
Nishit Anand

UNIVERSITY OF MARYLAND
FEARLESSLY FORWARD

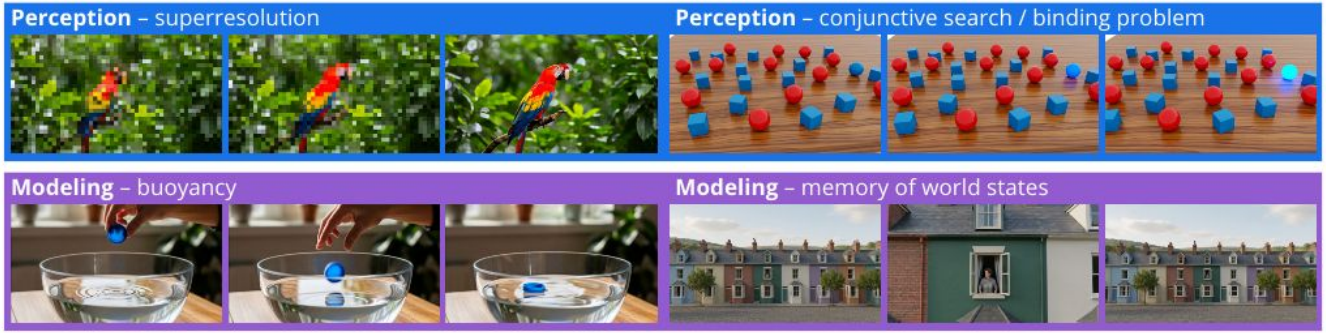
Introduction

The Illumination Problem

- Light *is* photography, and its essential for realism and perception
- Users want creative control over illumination in their images
- But illumination modelling is complex (surface property, geometry , reflections etc)



Learning Physics at Scale



Based on their perception of the visual world, diffusion models are starting to model it, including learning the physics of the real world

Figure: Video models are zero-shot learners and reasoners

The Open-Source Gap

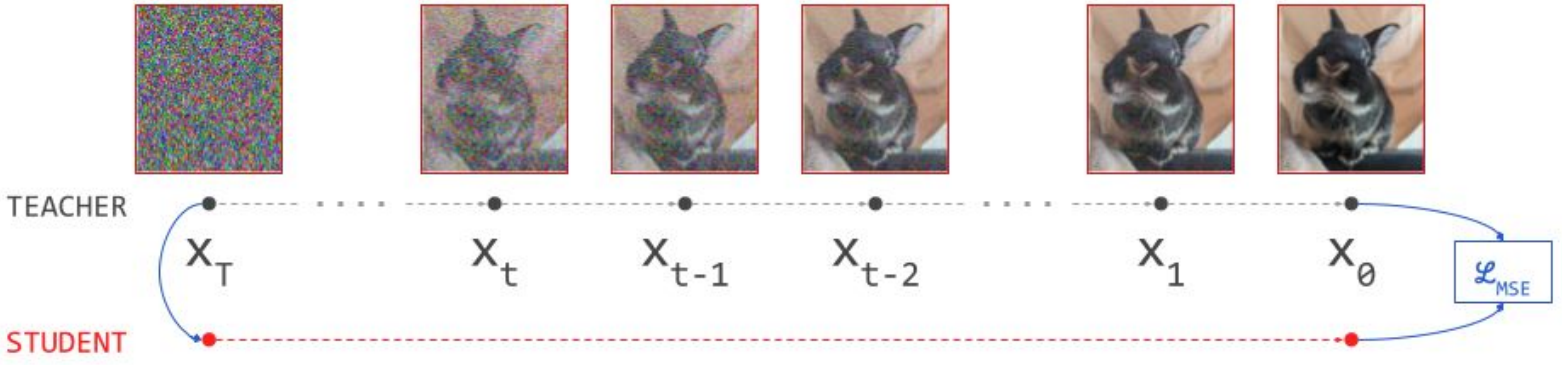
Qwen-Image 20B

Gemini Nano Banana

Open-weight and closed source models show illumination control.

But how can we achieve this with open-source models and data?

Distillation From Strong Models



Distillation from stronger teacher models is not optimal.

Risk of learning a narrow synthetic domain and still have a gap with closed models

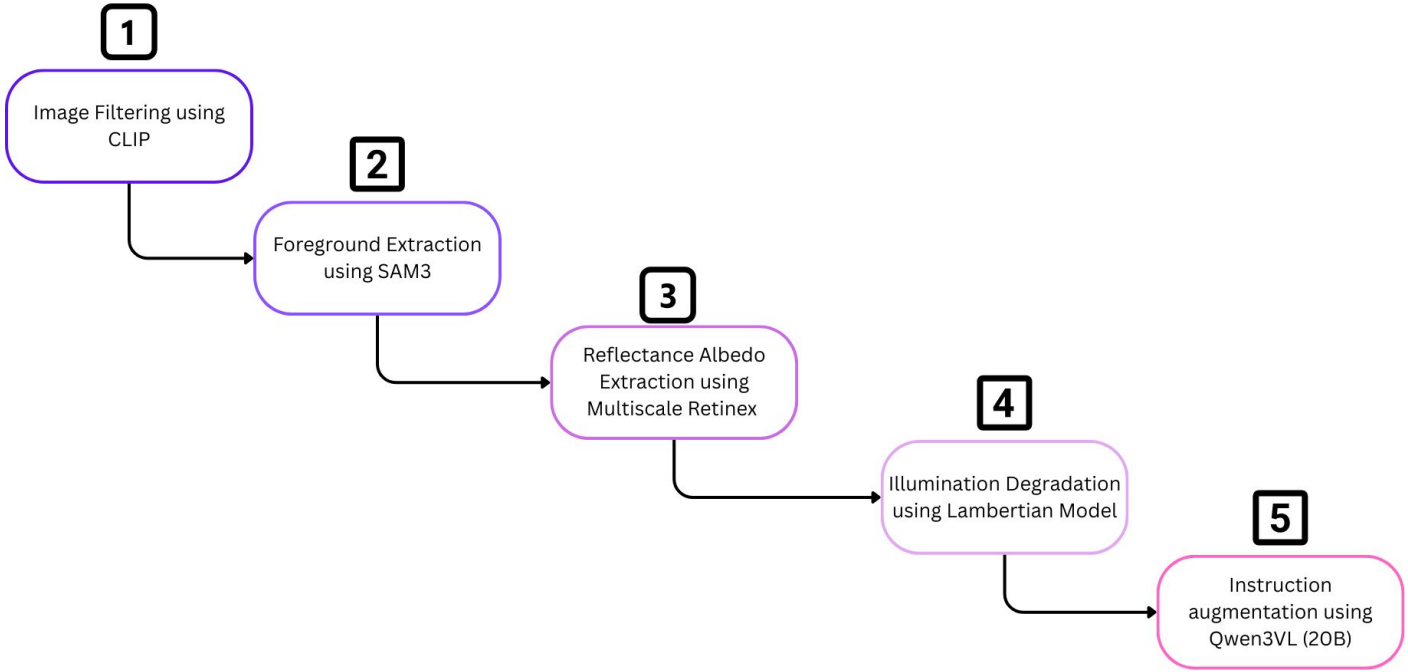
Mining Data In The Wild

- HQ images with illumination are abundant on the web.
- How to build a scalable open-source data engine to mine and annotate images in the wild for illumination modeling

Images from pexels.com

Data Engine

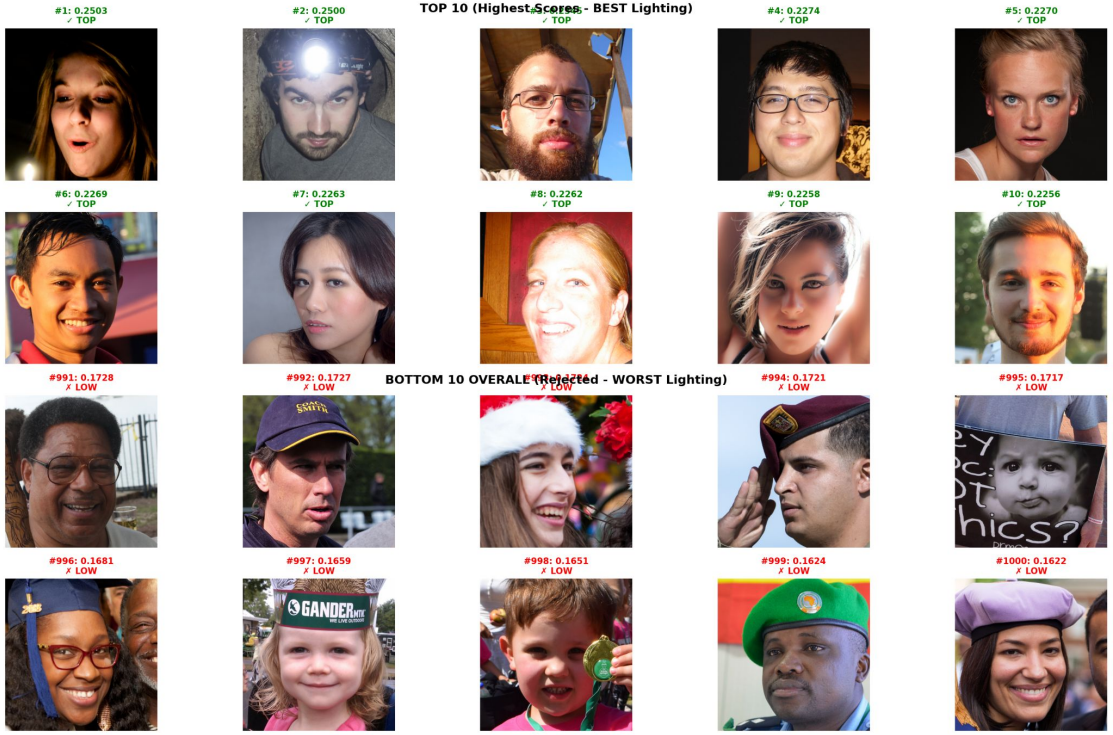
End-to-end Pipeline



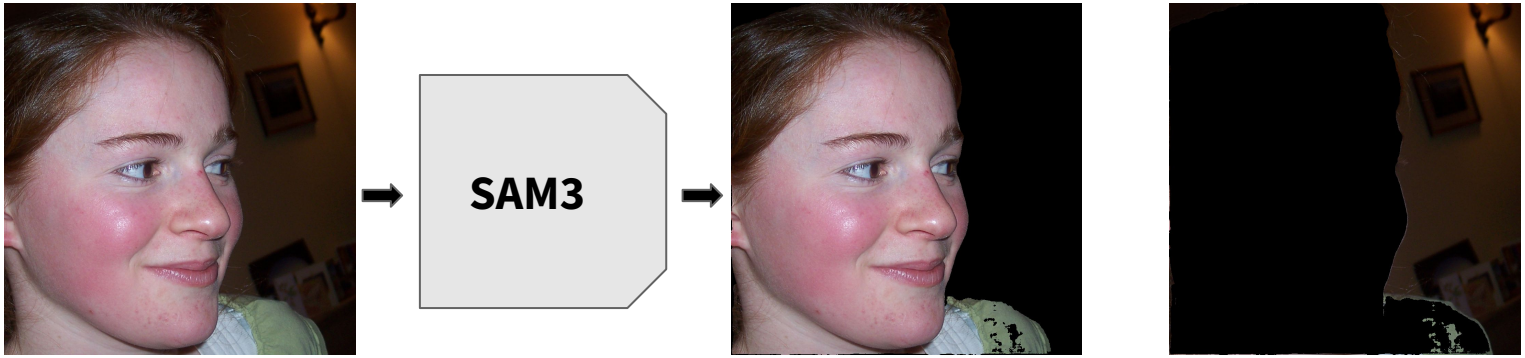
Initial Image Filtering

- Used 7 prompts
- Averaged their Similarity score:

1. "beautiful lighting",
2. "good lighting",
3. "well lit face",
4. "professional lighting",
5. "natural light",
6. "illumination",
7. "bright and clear lighting"

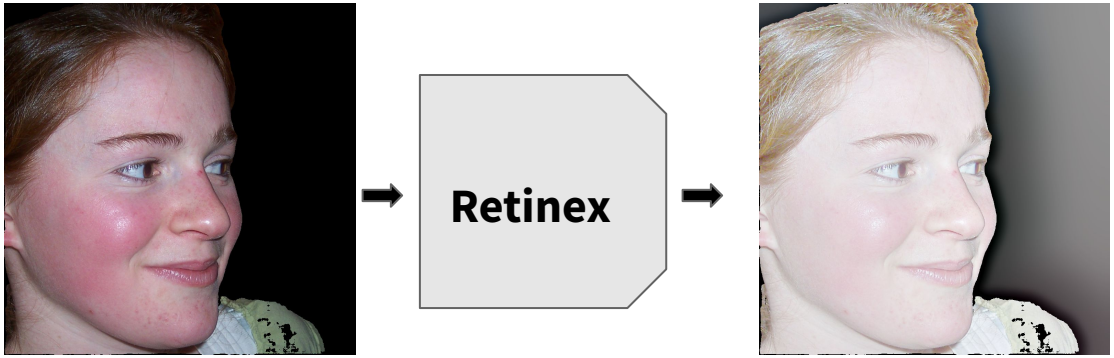


Subject Segmenation



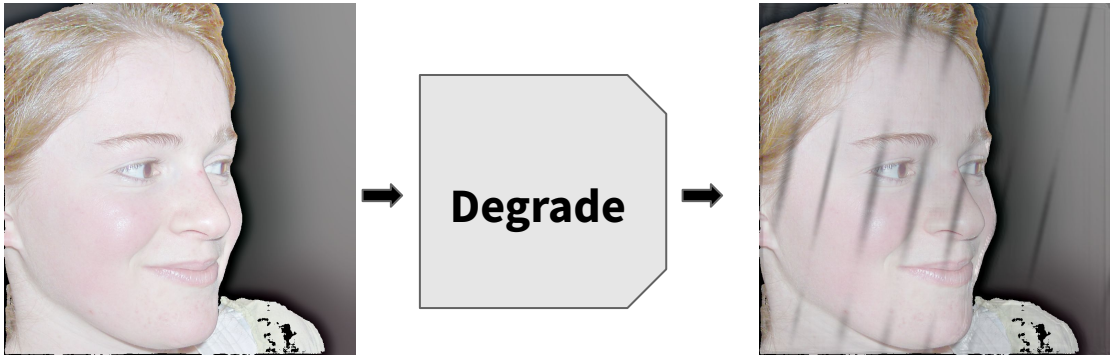
- Use SAM3 and a natural language prompt to segment the subject

Albedo Extraction



- Sample from a collection of albedo extraction methods to extract the albedo of subject

Further Degradation



- Apply random shadows to degrade images
- Use depth estimation for natural shadow placement
- Source high-quality shadow overlays from open stock image libraries

Soft Shaded Shadows

S3D – Soft Shading (Lambertian Model)

1. **Estimate surface normals** from the depth map.
2. **Sample a random light direction**, uniformly over the hemisphere.
3. **Compute Lambertian shading**
    - $I = \text{Ambient} + K_d \cdot (N \cdot L) I$
    -
    - **Ambient term:** 60–85% (keeps shadows subtle)
    - **Diffuse term:** dot product of normal and light direction
4. **Add procedural shadow patterns (from a library of shadow objects)**
    - Apply at **35–60% opacity** for soft, unobtrusive shadows.

Image Editing Instruction Generation

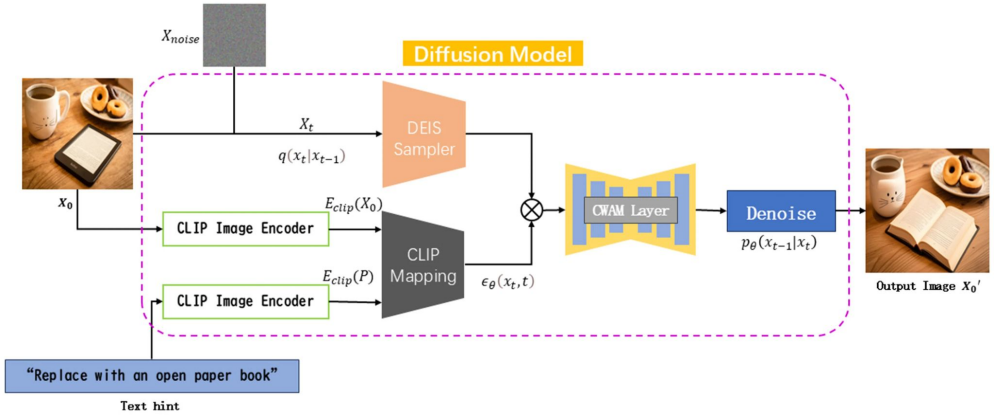
Qwen

Cinematic portrait with hard directional light from window blinds casting striped shadows across face, high contrast chiaroscuro lighting in dark interior space.

- Limiting to only paired image-text datasets will be restrictive
- We find that large vision language are accurate in their description of light sources and their effects
- We use Qwen3-VL 30B for generating captions

Training Details

- **Base Model:** Stable Diffusion 1.5
- **Learning Rate:** 1e-5
- **Epochs:** 250
- **Training Images:** 10,000
- **Test Images:** 1,000
- **Resolution:** 512×512
- **Training:**
- VAE and Text Encoder remain frozen, U-Net is fully finetuned

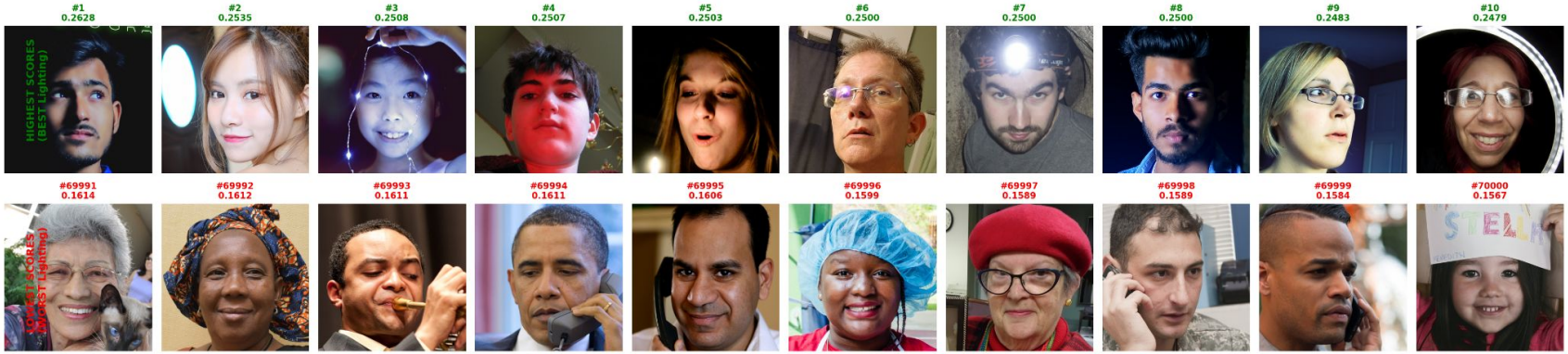


We use the Instruct Pix2Pix architecture

Results

Dataset Analysis

- Used ViT-B/32
- Manual verification to find threshold
- Selected images with similarity > 0.21
- 12000 images with score > 0.21
- 10K Train, 1K Test



Evaluation Metrics

- **LPIPS:** Perceptual similarity (learned features)
- **SSIM:** Structural similarity (luminance, contrast, structure)
- **CLIP Score:** Text-image alignment (does output match instruction?)
- **Identity score:** Face identity preservation (same person?)

Quantitative Results

- Generated images match through our fine tuned model are closer to ground truth image
- The identify of the person is preserved in the generated image

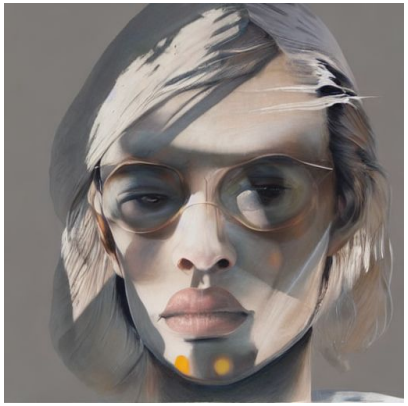
Metric	SD 1.5	SDXL	FLUX.1-dev	Our Model
LPIPS ↓	0.6346 ± 0.0901	0.6292 ± 0.0896	0.6504 ± 0.0787	$\mathbf{0.3002 \pm 0.0904}$
SSIM ↑	0.3802 ± 0.0951	0.4333 ± 0.1009	0.3726 ± 0.0974	$\mathbf{0.5667 \pm 0.1002}$
CLIP ↑	$\mathbf{0.2601 \pm 0.0280}$	0.2567 ± 0.0291	0.2520 ± 0.0303	0.2504 ± 0.0314
Identity Score ↑	0.0712 ± 0.0788	0.1088 ± 0.0980	0.0437 ± 0.0796	$\mathbf{0.7591 \pm 0.1823}$

Qualitative Results

Input Image

Instruction:
Harsh midday sun from directly above casts strong shadows under the eyes and nose, creating a high-contrast, naturalistic look with bright highlights on the forehead and cheekbones, suggesting an outdoor setting under a clear sky

Our model's output image



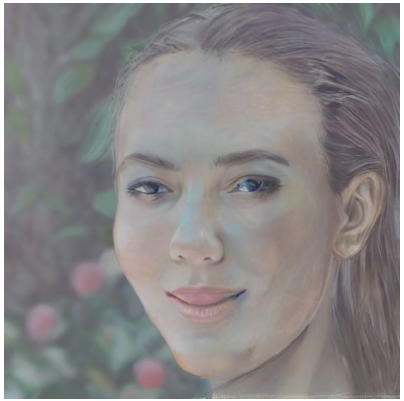
Baseline Stabe Diffusion 1.5 output image

Qualitative Results

Input Image

Instruction:
Soft, natural daylight illuminates the subject from the front-left, creating a gentle, even glow that highlights her features with a warm, flattering tone, while the blurred green foliage in the background suggests a serene outdoor setting.

Our model's output image



Baseline Stabe Diffusion 1.5 output image

Future Work

- We are currently using simpler albedo extraction approaches, we plan to implement recent state-of-the-art models.
- In this work, we limit to people faces domain, we plan to expand this broader categories