PROJECT DECISION LOG

Inter IIT Tech Meet 2025

BRIEF SUMMARY

This document captures the critical engineering decisions for the mobile-optimized **AI Driven Photo Editing App**. The system targets four core capabilities: **Text-Based Segmentation**, **Image Relighting**, **Object Removal**, and **Object/Background Replacement**.

CORE ACHIEVEMENT

Successfully transitioned from server-grade SOTA models (SAM3, DPT, SD-XL) to hybrid, CPU-compatible architectures (MobileSAM, MiDaS-small, LaMa-Dilated, SD v1.5) optimized for local mobile deployment.

1 MODULE: TEXT-BASED SEGMENTATION

Objective: Achieve on-device text-guided segmentation with latency under 15 seconds.

DL-01 Evaluating SAM3 for Deployment [REJECTED]

CONTEXT

Baseline model was Vanilla SAM3, selected for SOTA performance.

TECHNICAL ANALYSIS

Mobile devices lack the GPU power required for SAM3.

- **Baseline Latency (CPU):** 2.47 minutes (148s) – Infeasible.
- **Quantization (INT8):** Reduced to 129s – Still too slow.

Final Verdict: Abandon SAM3. Architectural overhead is too high for mobile CPUs.

DL-02 Hybrid CLIP + MobileSAM [PIVOT]

CONTEXT

Pivoted to **CLIP** (Text Encoder) + **MobileSAM** (Mask Generator).

BOTTLENECK DISCOVERY

Accuracy was preserved, but inference spiked to **76 seconds**.

- **Cause:** CLIP had to evaluate *every* candidate mask against the text embedding.

Final Verdict: Adopt Architecture, Optimization Required. CLIP evaluation is the new critical path.

DL-03 Prompt-Guided Search Reduction [OPTIMIZED]

INNOVATION

Introduced **Prompt-Aware Mask Filtering**. Use prompt semantics (e.g., "person on the left") to discard irrelevant masks spatially *before* embedding generation.

RESULTS

- Latency dropped from 76s $\rightarrow$ **27s**.

Final Verdict: Integrated. First configuration to achieve CPU viability.

DL-04 Deploying Tiny-CLIP [DEPLOYED]

STRATEGY

Replaced standard CLIP with **Tiny-CLIP** to minimize encoder latency.

PERFORMANCE MATRIX

- Tiny-CLIP + MobileSAM: $\approx$ 17s.
- **Tiny-CLIP + Pruning (Combined):** $\approx$ **13s**.

Final Verdict: Production Ready. Final shipping configuration.

Figure 1: Output by Final Architecture with an Average Latency of 13 sec

2 MODULE: RELIGHTING PIPELINE

Objective: Physics-accurate relighting with intuitive mobile controls.

DL-01 Depth Estimator Selection **[SELECTED]**

ANALYSIS

MiDaS-small selected over Large DPT. It provides the best balance of smoothness for shading and CPU/GPU-constrained inference speed.

Final Verdict: MiDaS-small. Sets the geometric foundation.

DL-03/04 Physics Core & Gesture Mapping **[ARCHITECTED]**

STRATEGY

Implemented an analytic shading engine (Lambertian + Phong) to capture basic physics.

UX DESIGN

Mapped swipes to directional light vectors and taps to point-light positions.

Final Verdict: Hybrid Approach. Physics provides the controllable baseline; User gestures map directly to 3D light vectors.

DL-06/09 Refinement Network & Domain Gap **[STRATEGY]**

MODEL

Tiny RelightNet: A lightweight network receiving RGB, Depth, Normals, and Physics-Shading.

TRAINING INNOVATION

Training on raw images failed. We corrected this by generating physics output on-the-fly during training to match the inference domain.

Final Verdict: Adopted. The network learns to *refine* physics rather than learn lighting from scratch.

DL-12 Final Pipeline Architecture **[FROZEN]**

SYSTEM FLOWCHART

`Input` $\rightarrow$ `MiDaS(Depth)` $\rightarrow$ `Normals` $\rightarrow$ `Physics Shading` $\rightarrow$ `Tiny RelightNet` $\rightarrow$ `Output`

Final Verdict: Deployed. Exported to ONNX (FP16/INT8) for NNAPI acceleration.

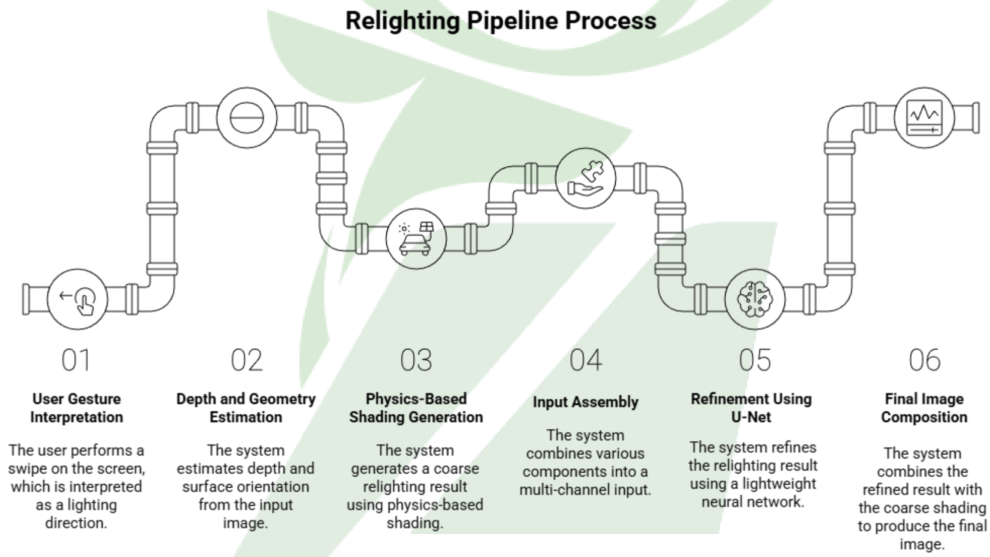

Figure 2: Final Architecture Adopted in the Project

3 MODULE: OBJECT REMOVAL

Objective: High-fidelity object removal with minimal footprint.

DL-01 Model Selection Benchmark [SELECTED]

CANDIDATES EVALUATED

- **LaMa-Dilated:** 0.54s latency. Sharp structure, best balance.
- **LaMa-Fourier:** 9s latency. Good quality, but too slow for interactive use.
- **Big LaMa:** 23s latency. High resource consumption, infeasible for mobile.
- **AOT-GAN:** 0.39s latency. Fast, but smeared textures.
- **MI-GAN:** 0.05s latency. Blurry blobs (Failure).

Final Verdict: LaMa-Dilated. Chosen for superior structural completion.

DL-02 Classical Approaches [REJECTED]

CONTEXT

Tested OpenCV methods (Telea, Navier-Stokes).

RESULT

Extremely fast but texture-less and blurry. Unsuitable for photo-editing quality standards.

Final Verdict: Dropped. Useful only for benchmarking.

DL-03 Quantization Strategy [MODIFIED]

EXPERIMENTS

- **FP16:** Size $\rightarrow$ 226 MB (39% drop). Negligible quality loss.
- **INT8:** Some degradations (color shifts, 43% filter error).

Final Verdict: FP16 Adopted. INT8 rejected due to visual artifacts.

DL-05 Final Pipeline Decision [DEPLOYED]

CONFIGURATION

`Mask (MobileSAM) → LaMa-Dilated (FP16) → Post-Processing`

Final Verdict: Production Ready. Compact, quantized, and stable across various mask shapes.

Figure 3: Output Comparison for FP16 Quantized Model

Figure 4: Output Comparison for INT8 Quantized Model

4 MODULE: OBJECT REPLACEMENT

Objective: Generative object replacement on CPU-only mobile hardware.

DL-01 Base Diffusion Model [SELECTED]

BENCHMARK

Compared **Stable Diffusion v1.5** vs. **DreamShaper**.

RESULT

SD v1.5 consistently ran faster across all setups while providing predictable results. DreamShaper had a slight speed penalty.

Final Verdict: Stable Diffusion v1.5. Selected for lower CPU latency.

DL-02 Step Budget Definition [CONSTRAINED]

ANALYSIS

Full quality requires 20-50 steps, which is infeasible on CPU.

- **5 Steps:** $\approx$ 32–43 seconds.
- **20 Steps:** $\approx$ 103–127 seconds.

Final Verdict: 5-Step Budget. Established as the practical ceiling for mobile CPU inference.

DL-03 Scheduler Selection [SELECTED]

COMPARISON

Tested Default, Euler_a, DDIM, and LCM.

IMPROVEMENT

While DDIM was effective, we subsequently integrated the **LCM (Latent Consistency Model) Scheduler**. This advanced scheduler is designed to produce high-fidelity images in as few as 4-8 steps, significantly reducing latency while withstanding accuracy drop-offs common in low-step regimes.

Final Verdict: LCM Scheduler Adopted. It offered the best trade-off for CPU-based inference.

DL-04/05 Final Configuration & Limitations **[PROTOTYPE]**

CONFIGURATION

SD v1.5 + LCM + 5 Steps.

STATUS

This configuration brings generation time within an acceptable window on CPU. However, complex prompts may still require more time.

Final Verdict: Good Enough for Prototype. Recognized as CPU-heavy; suitable for proof-of-concept but requires NPU/GPU for production refinement.

Figure 5: Outputs From Various Diffusion Models for the Prompt "A baby Panda"