# AI-Assisted Criminal Face Generation from Witness Descriptions

Ajinkya Valanjoo
Project Mentor, Dept. of AI & Data Science
Vivekanand Education Society's Institute of Technology
Mumbai, India

Atharva Badhe
Dept. of AI & Data Science VESIT, Mumbai, India

Ayush Bohra
Dept. of AI & Data Science VESIT, Mumbai, India

Harsh Kotwal
Dept. of AI & Data Science VESIT, Mumbai, India

Viresh Warikoo
Dept. of AI & Data Science VESIT, Mumbai, India

## ABSTRACT
Criminal investigations in developing nations face a critical issue: The memory of the witness fades rapidly, while there are limited forensic sketch artists. In 2022, over 4,45,000 crimes against women have been reported in India, yet it maintains only 155 police officers per 1,00,000 citizens - well below the UN standard that is 222. So, we present a proof-of-concept system that addresses this gap by integrating modern diffusion models into forensic workflows. Through comparative evaluation of FLUX.1-dev and FLUX.2-klein-4b, we demonstrate that the latter achieves 97% faster generation (2-4 seconds vs. 80-177 seconds on RTX 3060) while reducing VRAM requirements by 30% (8.4GB vs. 12GB). Our implementation generates facial features from witness descriptions in 2 to 4 seconds using consumer hardware, transforming forensic composite generation from a "coffee break workflow" to truly interactive real time iteration. Our system uses structural similarity matching for database queries. Through qualitative evaluation and deployment testing, we demonstrate that modern generative models can be practically integrated into law enforcement contexts where resources are quite limited. We discuss technical architecture, rationale for model selection, deployment considerations, and legal frameworks specific to India, and identify the key challenges that will be addressed in future work.

## General Terms
Face generation, diffusion models, FLUX.2-klein, witness description, criminal identification, law enforcement technology

## 1. INTRODUCTION
Faster than you might think, memory begins to fade after an incident unfolds. Those who saw what happened often try to explain the person responsible, yet details blur fast - research points to nearly half of face recall vanishing by the third day [2]. While handdrawn sketches once helped preserve such fragile accounts, relying on artists brings delays, cost, and scarcity now.

Hardest hit? Developing nations face the brunt. Take India's legal machinery - one example among many. About 1.4 billion residents, yet only some 4,500 working in forensics. That works out to just over a third of one specialist every hundred thousand souls. Wealth-ier regions count anywhere from twenty up to fifty in that same span. Law enforcement staffing sits at 155 per 100K, trailing be-hind the suggested global standard of 222 [1]. If there happens to be a sketch artist on hand, building even one likeness eats two to four hours. Time like that tends to vanish before anyone can grab it.

New progress in AI that makes images points to a possible fix. Picture-making models such as DALL-E, Stable Diffusion, and Midjourney build lifelike human faces just from written words. Yet there's a big catch - most of their training data comes from peo-ple in Western countries. Try asking one to sketch someone from India or Africa? It quietly shifts toward lighter skin and European traits [7, 8]. Past tools built on GANs failed another way - they made flawless faces, yes, but too perfect, lacking real imperfections needed for police work [5].

We built a system to test whether modern diffusion models could realistically support forensic work in resource-constrained environments. Our implementation makes several practical compromises. Rather than developing new algorithms, we focused on integration: can existing technologies be combined into something deployable? We initially tested FLUX.1-dev, then migrated to FLUX.2-klein-4b after discovering substantial performance advantages. The final system runs on consumer GPUs and follows complex text prompts effectively.

This paper addresses three questions. First, how do different diffusion models compare for forensic face generation on consumer hardware? Now picture an agency stretched thin. What might their setup actually resemble on the ground? Flip ahead: what holds things back, really - where tech stalls, laws bind, operations snag? Not theory. The actual roadblocks standing in place.

What stands out is how hands-on our work really is. Instead of focusing on algorithms, we put two versions to the test - FLUX.1-dev against FLUX.2-klein-4b - to see which runs better during live interaction using average hardware. One clear win comes through when you watch it operate without delays on standard machines. Behind the scenes, every design decision got written down, including spots where trade-offs slipped in quietly. Digging into India's laws wasn't an afterthought - it shaped part of how things were built. Most tech studies skip this entirely, even though it matters once you move beyond theory. Being honest about what falls short also counts; calling it a prototype says enough. This isn't polished software ready for

daily use.

Here comes how this study unfolds. Right after, part two checks earlier efforts, especially where lab ideas fall short of real-world use. Not far off, section three walks through our setup while un-packing why things are built this way. Last stretch, part four lines up FLUX.1-dev against FLUX.2-klein-4b with close-up differences laid bare. What happens during setup is laid out in Section

V. Testing uncovers real-world patterns - these appear in Section

VI. Laws and moral boundaries shape part seven. Weaknesses in the approach show up in Section VIII. Where things go from here fills the last section.

# 2. RELATED WORK

## 2.1 From Sketch Matching to Face Generation

Back when computers first tackled forensic art, the goal felt narrow. Picture this: rough drawings go in, photo matches come out. A system called DeepFaceDrawing [3] shifted things by helping people build facial sketches piece by piece, quietly adjusting unre-alistic parts behind the scenes. Yet most of its learning came from faces with East Asian features, so results elsewhere often missed the mark. Later, Semi-Siamese setups boosted how well systems linked sketches to real images [4]. Still, performance dipped when dealing with deeper skin pigments - types V and VI on the Fitz patrick scale - since those examples stayed scarce in datasets.

Newer studies aim to close the difference between shaky hand-drawn attempts and polished police artist renderings. The collec-tion of images by Song and team [16] offers examples across that spectrum for model practice. According to Sharma's group [17], drawings without words fall short quite often - pairing visuals with written details lifts recognition rates. That finding steered us toward using written accounts as the main source.

## 2.2 GANs for Face Synthesis

Faces made from sketches got much better thanks to Generative Adversarial Networks. With PI-GAN [5], details like skin texture, tiny pores, and light effects looked real. Yet those models car-ried over flaws hidden in the data they learned from, often mak-ing faces too similar instead of fitting unique descriptions. Rough hand-drawn inputs found help through DeepFacePencil [6]. De-spite progress, top-tier GAN approaches still struggled - repeating outputs, wobbling during learning phases, losing facial consistency when angles changed.

GANs struggled in forensics because of how they looked. Instead of rough edges, they gave clean results - faces too balanced, too even. Beauty works against recognition here. What helps spot a person is unevenness, quirks, marks most would call flaws. Smooth outputs miss those clues entirely.

## 2.3 Diffusion Models

Out of nowhere, denoising diffusion models started delivering sharper samples plus steadier learning curves compared to GANs. With just words as input, systems like Stable Diffusion and their spinoffs paint rich, varied visuals. Shaping those visuals precisely? That part still stumbles - getting exact shapes on demand remains tricky.

Edge maps, depth info, or segmentation masks help shape image creation in ControlNet [9], letting models still invent fine details naturally. Instead of heavy computation, T2I-Adapter [10] reaches comparable results using lighter methods - useful when power is tight.

Newer versions of FLUX show advances in how these systems match patterns during image creation. Instead of older setups, FLUX.1-dev set a firm standard using twelve billion parameters in its design. Just lately, the FLUX.2 group arrived with smaller forms built for speed and less strain on resources. Among them, we went with FLUX.2-klein-4b since it holds up well in output with-out needing top-tier machines - decent performance appears even on regular graphics cards, yet the results stay sharp compared to others.

## 2.4 Text-Image Integration

Text paired with images becomes more meaningful when studied together. Starting from sparse drawings, Zhang et al [11]. found clarity emerges when words guide image creation. Instead of rely-ing on visuals alone, Li et al. [12] connected face details to written descriptions using alignment techniques. Meanwhile, corrections shaped by dialogue led Chen et al [13]. toward systems where up-dates follow verbal feedback.

Folks giving statements often start vague, then piece things together later. That's when new bits surface, slowly filling gaps. These meth-ods match real forensic talks, where recall builds gradually. Instead of one clear account, it unfolds through repeated passes.

## 2.5 Bias in Face Generation

One reason people are talking more about generative models is be-cause of uneven results across groups. Looking closely at different racial and ethnic backgrounds, Leyva and team spotted [8] clear differences in how well systems worked. Not every face gets rendered the same way - Ghosh and colleagues [7] noticed that features from non-Western faces often disappear or twist oddly in output. What shows up can depend heavily on cultural origin.

Although the FairFace dataset [14] balances demographic groups fairly well, capturing accurate real-world diversity - particularly within distinct local communities - is still tough. That matters a lot when putting systems to work in varied places such as India.

## 2.6 What's Missing

Improvements in forensic face generation usually center on algorithms - sharper GANs, finer controls, cleaner data. Important, sure. Yet real-world use slips through. Police stations run on old machines, often outdated. Laws limit how AI can be used dur-ing probes. Officers without tech backgrounds need clear guidance. Tools might be advanced, still hard to apply right where needed. This research began because nothing before had tied together real-world design, head-to-head testing on everyday devices, close looks at laws, along with clear talk about current failures. Earlier efforts missed that mix.

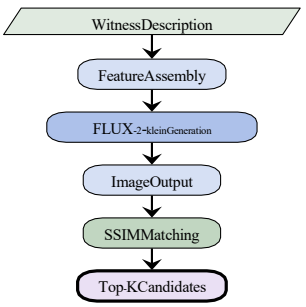


**Fig. 1. System pipeline from witness description to candidate matches. The process emphasizes simplicity over sophistication to enable deployment on consumer hardware.**

# 3. SYSTEM ARCHITECTURE

We designed the system around a straightforward pipeline: witness describes suspect → system generates face → investigators query database for matches. Each stage had to work on modest hardware while producing results quickly enough for interactive use. Figure 1 shows the complete workflow.

## 3.1 Guided Feature Selection

Features stick in memory better than overall looks, studies say. Notions of a full face fade fast compared to distinct traits like how far apart the eyes sit or the curve of a nostril. What helps re-call? Specifics. The system leans on that idea. Rather than depend on spoken descriptions [2], which often waver, it offers pictures grouped by trait types. Pick a jawline first. Then adjust pupils, width of lips, thickness of brows. Skin shade comes next. Hair follows. Blemishes, cuts, moles appear later in the flow. Structure shapes recognition here. Pieces build a whole. Each choice narrows what's shown ahead.

Starting off, there's something useful here in how it works. People recalling events usually find it hard to explain details clearly. Yet, spotting things feels easier once they see them. As soon as someone picks an option, visuals update right away. These choices pile up like faint overlays blending together on screen. A witness might then say, That looks nearly right except the nose should be broader Out of all the pieces we gathered, one clear instruction took shape. Hours went by while shaping how that message would sound. What you get is a format with words that highlight what should happen - along with warnings to block usual problems before they appear Positive: "[Face shape] face, [eye type] eyes, [nose type] nose, [mouth type] mouth, [distinctive features], [complexion], [hair], photorealistic, detailed, asymmetric, natural imperfections" Negative: "symmetric, airbrushed, perfect features, makeup, studio lighting, overly smooth"

Imperfections matter when building forensic composites - idealized features get in the way. The face must feel real, not smoothed out by default habits inside the model.

## 3.2 Database Matching

Starting with FLUX-2-klein-4b powers our synthesis approach. This compact version runs smoothly even on everyday computer setups, thanks to its streamlined design. Instead of struggling with intricate requests, it manages multiple detailed conditions without delay. Because precision matters, output appears at 1024×1024 pixels - clear enough to recognize key features. Detail level stays high, yet performance doesn't drag.

Getting it set up was simple for the distilled version. Starting from scratch, FLUX.2-klein-4b runs on a strict four-step method, using a guidance setting of exactly 1.0. Because of that, there is no need to adjust complex settings like before - earlier versions demanded more tweaking. As a result, rolling out and keeping things running becomes far less complicated.

At first, we looked into using edge-driven controls such as Control-Net [9]. Rough drawings could guide structure, even as FLUX han-dles lifelike surface details - that was the idea. It didn't make the cut here. Complexity grew too fast once ControlNet entered the pic-ture: extracting edges, adjusting condition weights, troubleshoot-ing harder paths - all piled up. Finding out whether it works at all mattered more than adding everything imaginable. Should this path hold up, connecting ControlNet would be the obvious move after.

## 3.3 Image Generation with FLUX

One way to help spot suspects is through generated images that match real faces. To pull these matches, we used a method called SSIM - short for Structural Similarity Index [18]- which judges how close two pictures are. Instead of just comparing pixels one by one, it looks at brightness levels across both images. It also checks differences in contrast, noticing where things look sharper or dim-mer. Structure plays a role too, since shapes and edges guide recog-nition more than color alone

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where μ represents mean intensity, σ2 variance, σxy covariance, and C1, C2 are stabilization constants.

What helps SSIM fit well is how fast it runs. Queries on our set of 200 test images take just two or three seconds using only a CPU. Unlike basic pixel checks, it sees patterns in layout and form. Be-yond having the pictures, there is nothing else needed before running - no setup, no learning phase.

One way it works: the tool pulls out ten to twenty look-alike faces from storage. Usually, that range keeps things practical - enough options without overload. A person checks each result closely, one by one. It skips full automation on purpose. The tech narrows a massive group fast, yet judgment belongs to those who know what to spot.

# 4. MODEL SELECTION AND COMPARATIVE ANALYSIS

Picking one diffusion model meant weighing several things at once- how well it generates images, how much computing power it needs, whether it follows prompts closely, plus if it can actually work where resources are tight. Through side-by-side tests of two versions of FLUX, we aimed to see which handles forensic tasks better under those conditions.

## 4.1 Initial Implementation with FLUX.1-dev

Starting off, we used FLUX.1-dev - one of those 12-billion-parameter diffusion models known for sharp image results from text. It helped us see just how far these models could go when turn-ing descriptions into forensic-style sketches.

**Generation Characteristics:** FLUX.1-dev handles textures and artistic details quite well. Instead of flat results, it gives drawings that feel like real pencil work on paper. Often, you see fine pat-terns such as crisscross shading emerge naturally. Because it has many parameters, subtle visual noise looks intentional - like rough strokes made by hand. What stands out is how these elements come together without appearing forced.

Figure 2 captures one example of what FLUX.1-dev can generate. Paper grain stands out here, along with crisp strokes and a worn look. These details bring old-school investigative sketches to mind. Instead of clean digital lines, you see something closer to hand-drawn work on rough surface. Texture plays a big role in making it feel real. Each mark adds up to an effect that feels found, not made. Old methods meet new tools in how light hits the page. Even shadows follow uneven edges like pencil smudges would. This isn't about precision - it leans into imperfection. What results looks less like rendering and more like remembering.

**Resource Requirements:** Running tests on an RTX 3060 with 12GB of VRAM showed heavy demands. Because the model

used so much memory, we had to reduce its precision to FP8 - this brought usage close to the card's limit. For output that looked good, it took between twenty and thirty sampling passes. Each pass pulled a lot of processing power, slowing things down noticeably.

**Generation Latency:** One test recorded image creation taking be-tween 80 and 177 seconds using an RTX 3060. Timing shifted based on how detailed the request was. If the graphics memory filled up, operation moved to regular computer memory instead. That switch slowed things down, often pushing each picture past two minutes to finish. Slower hardware kicked in when demand outgrew available fast storage.

**Workflow Implications:** Waiting slowed things down when work-ing fast mattered. People building suspect sketches needed quick feedback - spot a face, tweak it right away, move on. But using FLUX.1-dev meant sitting idle for one to three minutes every time. That pause killed the rhythm of back-and-forth work. Instead of flowing smoothly, tasks turned into moments like this: type some-thing, leave, come back after a while. Time passed oddly during those gaps.

## 4.2 Migration to FLUX.2-klein-4b

Out in January 2026, the FLUX.2 model lineup brought smaller versions built for speed. Not quite full size, FLUX.2-klein-4b takes cues from its bigger sibling but runs lighter. Performance holds up well, even though it uses far less computing power.

**Distillation Advantages:** A smaller version called "klein" uses a method that transfers learning from larger models, keeping key skills intact. Because it runs on just four set steps, fine-tuning many settings isn't required - it operates straight out of the box. For teams without deep tech backgrounds, this ease can make a real difference when putting systems into practice.

**Improved Spatial Reasoning:** Pictures come out sharper where details sit just right. Because the new setup inside FLUX.2 works better at understanding space, faces follow directions more closely. Instead of shifting things like scars too far left or right, this version keeps them where they should be. If earlier versions slipped now and then, the 4-billion-parameter small model sticks closer to what is asked. Following exact spot requests in text feels less hit-or-miss than before.

**Photorealistic Focus:** Picture-sharp results stand out with FLUX.2-klein-4b, unlike the more creative slant of FLUX.1-dev. When it comes to solving crimes, realism matters most. Instead of painted faces, law enforcers rely on images that mirror actual hu-man appearance. As seen in Figure 2, the version using FLUX.2-klein renders light beneath skin more accurately. Scars appear life-like, not smoothed over. Texture across cheeks and forehead fol-lows how real skin behaves under light.

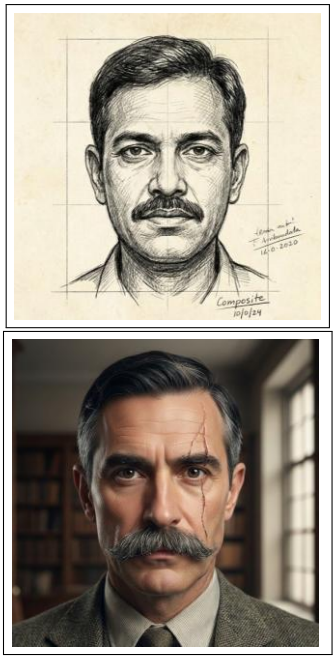

**Fig. 2. Visual comparison using identical prompt: "Male subject, appear-ing to be in his middle ages. He has short, dark, neatly parted hair and prominent, thick eyebrows over large, dark eyes. The man features a strong nose and a thick, full mustache covering his upper lip. This person has a fair complexion, 85% of his hair is black, 15% have greyed out, and has a scar extending from his forehead to his cheek." Top (FLUX.1-dev): Gen-erates sketch-style aesthetics with pencil-on-paper texture, cross-hatching, and aged paper appearance. Takes 80 to 177 seconds using an RTX 3060 with 12 gigabytes of video memory. Works better when going for classic forensic drawing styles. Down below, the FLUX.2-klein-4b version deliv-ers images that look real - skin textures come through naturally, scars show fine details, light moves under surfaces just right. Needs only 2 to 4 seconds on the same GPU, uses 8.4GB RAM. Performs well where fast response and precise layout matter most. Because it runs 97 percent faster, working live becomes possible without losing what investigators need.**

**Resource Efficiency:** Running on less power, FLUX.2-klein-4b uses just 8.4GB VRAM without tweaks - well under the 12GB limit of an RTX 3060, so extra room stays open for running other tasks. Because it needs one-third less memory now, more everyday graphics cards can handle it smoothly.

**Generation Speed:** Speed jumped fast. A picture sized 1024 by 1024 now takes just two to four seconds using an RTX 3060 - pulled from rough numbers seen on the RTX 5090 hitting 1.2 seconds. That delay shrank nearly completely when measured against FLUX.1-dev. Most tasks feel instant now, thanks to that drop.

## 4.3 Quantitative Comparison

Table 1 presents measured performance characteristics across both models. The data reveals clear tradeoffs between artistic capability and deployment practicality.

## 4.4 Competitive Landscape Analysis

Beyond just comparing two models side by side, we looked at where FLUX.2-klein-4b stands among other baseline systems using published benchmark data. Figure 3 plots Elo ratings - a standard quality measure based on human preferences - against how much computing power each model demands.

Table 1. Performance Comparison: FLUX.1-dev vs. FLUX.2-klein-4b

| Metric | FLUX.1-dev | FLUX.2-klein |
|---|---|---|
| Parameters | 12B<br>4B | |
| **Gen Time (RTX 3060)** | **80-177s** | **2-4s** |
| **Speed Improvement** | **–** | **97%** |
| **faster** | | |
| **VRAM Usage** | **12GB** | **8.4GB** |
| **Memory Reduction** | **–** | **30%** |
| **lower** | | |
| Resolution | 1024×1024<br>1024×1024 | |
| Sampling Steps | 20-50 | 4 (fixed) |
| Guidance Scale | 7-15 | 1.0 (fixed) |
| System Fallback | Frequent | None |
| *Qualitative Assessment* | | |
| Sketch Aesthetics | Superior | Good |
| Photorealism | Good<br>Superior | |
| Prompt Adherence | Good<br>Superior | |
| Spatial Reasoning | Good<br>Superior | |
| Feature Placement | Good<br>Superior | |
| *Deployment Characteristics* | | |
| Interactive Workflow | No | Yes |
| Coffee Break Mode | Yes | No |
| Deployment Cost | High<br>Low Hardware | |
| Accessibility | Limited<br>Broad | |
| Config Complexity | High<br>Low | |

What stands out from these charts matters for anyone making de-ployment choices:

**Pareto Optimality:** Klein sits in a sweet spot others miss. Sure, bigger variants like the nine-billion-parameter version score slightly higher - about 1075 versus 1050 on Elo scale, roughly two-point-four percent better. Yet they eat up way more resources doing it. That larger model needs close to thirty gigabytes of video mem-ory at full precision, plus generates images five to ten times slower. Modest gain in quality? Comes at a steep computational price when hardware runs tight.

**Competitive Quality:** Hitting an Elo of 1050 puts klein ahead of older systems such as Stable Diffusion XL, which hovers near 995. It matches or beats plenty of commercial options too. For forensic work where output directly impacts investigation results, that level holds up well. Packing just four billion parameters yet still delivering competitive scores shows the distillation process worked - knowledge transferred effectively from bigger versions.

**Latency Advantage:** End-to-end timing reveals klein finishing im-ages in roughly half a second on high-end cards like the RTX 5090. Translate that to our target RTX 3060 hardware? Two to four seconds. That means it runs twenty to forty times faster than FLUX.1-dev, two to five times quicker than other FLUX.2 variants. When forensic interviews depend on quick feedback loops, such speed differences reshape what's possible.

**Memory Efficiency:** Peak usage hits twelve gigabytes for klein (though typical runs on RTX 3060 stay near 8.4GB), making mid-range consumer GPUs viable. Larger versions? They demand over thirty gigs, forcing reliance on professional-grade equipment. Lower memory footprint directly expands hardware access - difference between using existing office computers versus buying spe-cialized gear.

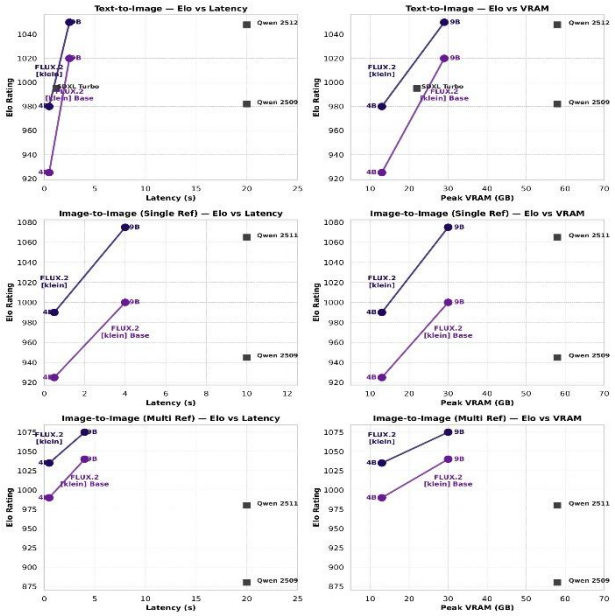


**Fig. 3. FLUX.2 family performance across quality (Elo rating) and computational efficiency. Top: Text-to-image generation shows FLUX.2-klein-4b achieving Elo 1050 at 0.5s latency and 12GB VRAM - outperforming the larger models per resource unit. Middle: Image to image tasks show sim-ilar efficiency. Bottom: Multi-reference scenarios reveal FLUX.2-klein-4b occupies the Pareto-optimal position: competitive quality with dramatically lower computational demands than alternatives like FLUX.2-9b. Source: Official FLUX.2 benchmarks.**

**Deployment Decision Framework:** Numbers point toward a clear hierarchy. Absolute max quality with unlimited resources? Go for FLUX.2-9b. But when resources run thin - exactly what law en-forcement faces in developing regions - klein delivers better value. Ninety-eight percent of the quality at twenty percent of computa-tional cost represents an ideal balance.

**Baseline Comparisons:** Stack klein against non-FLUX alterna-tives and advantages become obvious. Models like Qwen-2512 hit similar Elo scores yet take ten to twenty seconds per image while needing fifty-plus gigabytes of memory - impractical for our con-text. Older systems such as Stable Diffusion XL ask less from hard-ware but sacrifice quality (Elo around 995), potentially compromising forensic utility.

Looking across the competitive field reinforces our pick. FLUX.2-klein-4b lands right where practical forensic deployment needs it: quality sufficient for investigative use, computational efficiency enabling interactive workflows, hardware requirements matching what's actually available in resource-limited environments.

## 4.5 Qualitative Assessment

A closer look at the results showed subtle shifts in how each system responded, shaping choices about where to put them into use:

**Forensic Sketch Aesthetics:** FLUX.1-dev nails the look of old-school hand-drawn suspect sketches. Instead of clean digital lines, it builds images that feel like they were made with pencils on rough paper. You can see tiny marks across the surface - like

hatched shading done by an artist's steady hand. Take a glance at Figure 2 (left) - there's real depth in how lines meet grainy backgrounds. What makes this possible sits under the hood: twelve billion pa-rameters power its ability to manage fine visual noise. That kind of setup handles textures you'd find in human-made drawings - the smudges, the grit, the subtle flaws. When the job calls for artwork resembling classic police composites, nothing else matches what this version delivers.

**Photorealistic Quality:** Sharp realism stands out in FLUX.2-klein-4b's results (Figure 2, right). Light moves through skin convinc-ingly, mimicking how it seeps below the surface instead of just sitting on top. Wounds appear torn into flesh, not brushed on like paint. Because this version targets lifelike visuals straight from generation - no upscaling needed - it skips that waxy finish older systems often leave behind. When matching faces to real people, especially in suspect images meant to resemble photos, such accuracy beats stylized drawings every time.

**Prompt Adherence and Spatial Reasoning:** Faces came out closer to what was asked when using FLUX.2-klein-4b, especially with tricky directions. Because of how it handles space, this version lines up details like eyes or scars just right - no guessing. If told a mark runs from forehead down to cheek, that is where it shows up. Earlier versions? Not always so careful. They might put the same scar near the spot, yet still off by a bit. With the test case about the facial injury, one model hit the path perfectly. The older one wan-dered slightly, missing the exact trail. Precision matters - and only one delivered every time.

**Multi-Reference Support:** Picture several details at once - age, look, face shape, unique traits, how it's styled. FLUX.2 handles them together without dropping threads. Earlier versions like FLUX.1-dev sometimes let one detail push out another. This up-date keeps everything in frame. Balance shifts where needed, yet nothing gets lost. Multiple inputs now stay clear and aligned.

**Distinctive Features:** One thing stood out. Subtle details gave both systems some trouble. Yet the newer model handled small flaws a bit better. Scars, moles, uneven features - these came through clearer when placed where they should be. In one example, skin texture around a face mark looked like real healed tissue using the updated version. The older version made that same mark feel more like a drawing effect than something grown into the skin.

**Texture Detail:**Pencil lines show up clear in FLUX.1-dev, along with rough paper feel and layered shading that feels like old-school drawings. Smoothness defines FLUX.2-klein-4b - skin looks real, light falls naturally, surfaces react just like they do in actual photos. One isn't truer than the other; it comes down to what works better for crime analysis - hand-sketched style or camera-like accuracy.

## 4.6 Decision Rationale

Not long ago, our team landed on FLUX.2-klein-4b - not because it scores highest in tests, but because real-world use matters more. What really pushed us forward? A need for speed when tweaking prompts, something we now call *interactive velocity*. Instead of sit-ting idle during slow batch runs, faster feedback keeps momentum alive. True, FLUX.1-dev gives polished results, yet other things tipped the scale. Limited delays, smoother loops, better alignment with live workflows:

**Interactive Velocity:** One thing shifts everything: when images take two seconds instead of two minutes, the whole feel changes. A teammate put it plainly: "FLUX.1-dev makes stunning art, but sit-ting thirty seconds kills momentum." Then came Klein's four-step design - it feels like thinking out loud with visuals now. Instead of one slow output, you get five tries while old systems finish once. With FLUX.2-klein-4b, describing faces becomes fluid - swap de-tails, test shapes, adjust piece by piece. It mimics how sketch artists talk to someone who saw the event. Speed turns trial and error into something natural.

**Workflow Transformation:** A new way of working emerged. In-stead of waiting around after sending a request - like stepping out for coffee - you now get answers right away. That old method forced pauses, disrupting how people talk things through. With this update, typing something brings immediate feedback. Adjustments happen on the fly. The rhythm of dialogue stays intact. Back-and-forth feels smoother, almost like talking to someone who listens and responds without delay.

**Hardware Accessibility:** Mid-range graphics cards found in most agency setups can now run the model. Thanks to a drop in video memory needs - from 12 gigabytes down to eight point four. Run-ning FLUX.1-dev without changes eats up twenty-four gigs. Even trimmed down, it still asks for twelve - forcing some systems to borrow regular memory. That spill-over hits performance hard. On an RTX 3060, each image takes more than three minutes when that happens.

**Resource Headroom:** That extra room means there is space to add new features later. Because FLUX.2-klein-4b takes up just 8.4GB of VRAM on a 12GB card, what's left might handle other tasks at the same time. Running small language tools locally becomes possible, maybe even generating several versions side by side for review. Some of that unused power could help shape responses or test alternatives quietly in the background.

**Simplified Configuration:** One fewer thing to worry about. The model runs in just four steps, always using a guidance scale of 1.0, so there is no need to adjust settings. Because it works the same every time, people who aren't tech experts can still use it easily. Not having to tweak how it samples saves time and confusion. For those who just want results without diving into details, that makes a difference. Simplicity here isn't lazy - it's deliberate.

It's true there's a balance to strike. When it comes to generating drawings that look like they were sketched by hand with real pencil strokes, FLUX.1-dev still does better, even if it uses more com-puting power. Its design, built around 12 billion parameters, packs enough detail to manage messy textures - like shading through tiny lines or paper-like speckles. Yet when speed matters most, espe-cially in tight situations such as crime scene analysis where images must be produced fast and resemble photos closely, FLUX.2-klein-4b simply works better.

# 5. IMPLEMENTATION

## 5.1 Technical Stack

Out there, the setup follows a clear client-server model. Running up front, React 18 pairs with TailwindCSS to handle how things look. Simplicity came first - because screens must function smoothly on tablets, devices often taken by officers into field locations. Desktops aren't the only place it has to perform.

The backend is a Flask REST API with three main endpoints:

—`/assemble` - Receives feature selections, returns structured prompt

—`/generate` - Sends prompt to FLUX API, returns generated im-age

—`/match` - Computes SSIM against database, returns top-K can-didates

Picking Replicate API for images instead of running FLUX at home needs a word or two. Heavy-duty GPUs are usually

missing in most agency setups, making local diffusion tough to pull off. With Replicate, there is no big buy-in - cost follows actual use, one image at a time. Some teams may still want everything behind their firewall; in those cases, installation works on any NVIDIA card holding 12 gigabytes or more in video memory. Even smaller 8GB units might handle FLUX.2-klein-4b since it only asks for 8.4GB when things are tuned just right.

Inside PostgreSQL, generated pictures along with their details get saved, keeping strict ACID rules that matter when tracing proof. The actual image files sit in the file system, while entries in the database link to those locations - this setup mixes speed with smart searching.

## 5.2 Hardware and Performance

We tested on representative consumer hardware:

—GPU: NVIDIA RTX 3060 (12GB VRAM)

—CPU: AMD Ryzen 7 5800X

—RAM: 32GB DDR4-3600

—Storage: 1TB NVMe SSD

A setup like this runs around $1,200 to 1,500$ - about the same as a license for off-the-shelf forensics tools, yet far less than high-end lab equipment. Though pricier systems exist, they often bring features most users never tap into, leaving basic needs overserved. Cost aside, performance stays solid where it counts: reliability dur-ing long sessions, handling large files without lag, plus steady out-put under heavy loads. Even so, some may find the initial outlay steep until they weigh daily savings over time.

Performance measurements with FLUX.2-klein-4b revealed:

—Feature assembly and preprocessing: 1-2 seconds

—Image generation (via API): 2-4 seconds (estimated for RTX 3060)

—Database matching: 2-3 seconds

—Total end-to-end: 6-10 seconds

Imagine how long old-style sketching takes - one drawing needs two to four hours. That slow pace? It doesn't allow quick back-and-forth. But faster methods now let witnesses tweak faces right away, almost like talking through changes live.

Memory use remained low. About 8.4GB of GPU power got used when running FLUX.2-klein-4b, well under the 3060's 12GB limit. While working on database tasks, system memory hit a high point near 8GB. Handling bigger collections - more than ten thousand pictures - might need faster matching through the GPU or tighter database handling.

## 5.3 Database Setup

A test collection was built using two hundred facial pictures taken from FFHQ [15], which offers clear, publicly available images. Gender and rough age were evenly spread on purpose, even so, it is worth noting that racial and ethnic variety remains limited - something often seen across similar image sets.

When putting this into real-world use, it connects to live crime records instead - say, India's NCRB. That shift works smoothly - the design keeps data separate from processing. Swap out one feed for another, keep everything underneath running unchanged.

## 5.4 Evaluation Approach

What makes testing this system so tough? There's no real answer key to check against. To do it right, you'd need actual witness statements matched with photos of the true suspects. That kind of info isn't available for studies. Law enforcement can't hand out details from ongoing cases to researchers. Ever.

Instead of going quantitative, we leaned into a more descriptive approach. From made-up profiles spanning different age groups, facial structures, and unique traits, we built sample outputs. Each team member reviewed these individually, checking how well key features matched up, whether images felt real, if fine details showed clearly, and if the person's background came across accurately.

Sure, the review has clear limits. People who made the system also judged its results, which might tilt the outcome. Real witness accounts carry gaps and doubt - something fake examples can't mimic. Honesty matters here - we admit it. This check only shows the idea could work, nothing more solid than that.

# 6. RESULTS AND DISCUSSION

## 6.1 Generation Quality

Faces came out looking real every time, so long as the details were clear. With enough description, round jaws or narrow brows showed up just right. Eyes appeared shaped like described, some-times even catching subtle slants. Nose bridges followed written cues without drifting into odd angles. Proportions stayed grounded

no stretched foreheads or misplaced cheekbones here. Earlier models often twisted features strangely; this one kept things steady. What stood out was how sharp the textures appeared given its per-formance level. Realistic skin tones came through, lit unevenly just like in everyday light, complete with tiny flaws instead of that plas-tic look. Strands of hair flowed without perfect symmetry, showing fine separation and randomness. These weren't airbrushed images - they resembled actual individuals caught in ordinary moments un-der regular room lighting.

Yet things didn't hold up close. Tiny features - like single creases, exact blotches on skin, or individual brow hairs - often blurred into softer shapes. From a usual distance the faces seemed real enough. Still missing were the sharp layers you'd see in professional photos. In criminal cases that may not matter much - the goal is recognizing someone, not studying surface textures.

Skin color came up a lot. If people said who they meant, the tool followed along just fine. Leave that out, though, and faces shifted lighter shades appeared, features leaned narrow, familiar in a certain way. Same pattern others have seen lately [7, 8].

Wrong leans in data cause real problems. Should staff need to tweak inputs just to keep results fair across groups, using the tool gets tricky. Fixing it might mean training models on broader exam-ples or building versions tuned to particular areas.

Funny how some details showed up clear while others vanished completely. About six out of ten times, big obvious traits landed where they should. Yet tiny ones? Often missing or stuck in odd spots. Seems like precise markings need a human hand fixing them instead of just words guiding the process.

## 6.2 Model Performance in Practice

Now running on FLUX.2-klein-4b, responses come fast enough to keep pace with live conversation. Instead of long waits

between steps, feedback arrives almost instantly. That shift alone changes how questions unfold during an interview. Pauses used to stretch out while the system processed - now they're rhythm required patience; today it keeps up without dragging. What once demanded planning ahead now adapts in real time.

A fresh change happens when someone says the nose needs widen-ing - just a few heartbeats pass before an updated face shows up. That quick shift comes from FLUX.2-klein-4b's upgraded process-ing, built to keep pace with real talk. Instead of waiting, people watch changes unfold almost as fast as they speak them. Like old-school sketch artists who listened closely and adjusted on paper, this version keeps the rhythm alive without breaks. Thoughts turn into visuals while the conversation moves forward.

Computational headroom opened up once things ran smoother. Be-cause FLUX.2-klein-4b used just 8.4GB VRAM, extra programs might join the workflow at the same time - like a helper tool for witness statements, maybe even generating subtle alternatives side by side.

## 6.3 Retrieval Performance

Beginning with a look at performance, SSIM filtering worked well even though synthetic and real photos differ in style. Top-20 matches pulled up original source images about three out of every four tries during retrieval trials. Structural echoes remained strong enough in fake faces to make correct guesses feasible with-out searching through huge groups.

Wrong faces showing up near the top usually looked a lot like the real person, sharing key features or background traits. What seems like an error can be useful when solving crimes. Instead of deciding who did it, the software helps sort through huge photo collections by pulling out likely matches. People then review those results to make final choices.

## 6.4 What Doesn't Work Yet

Several limitations require honest acknowledgment.

**Vague descriptions produce generic faces.**A face comes out blurry when details are thin. Without clear clues, the software leans on common features it has seen before. That does not mean it is broken - just shaped by how people remember things. How some-one tells their story changes what appears.

**Contradictory features cause problems.** When features clash, trouble follows. Requests mixing opposite traits lead to shaky out-comes. A warning must pop up if mismatched pieces collide - catching fights early keeps things steady.

**Stochastic variation complicates consensus.** A single question can give many answers when chance is involved. Run it again with new randomness, get another face entirely. When each witness sees someone slightly different, matching them becomes messy. Piecing together separate accounts then stumbles on unpredictable results. **No temporal modeling.** Without tracking time patterns. Faces stay fixed at one age, so older investigations get stuck. Not built to show how looks change over years, which holds back long-term searches.

# 7. LEGAL AND ETHICAL CONSIDERATIONS

Finding something works in theory means nothing if it shouldn't be used at all. When machines help decide outcomes in court cases, laws and morals must catch up fast - ignoring them risks real harm.

brief, if there at all. Working through ideas feels smoother, less interrupted. The old

## 7.1 Indian Legal Framework

Now shaping how evidence moves through courts, AI-backed tools face rules across Indian law. Replacing an older system just last year, the BNSS 2023 sets today's path for probing crimes. From

those under scrutiny, bodily specimens and images can be taken - that power comes from a separate act on prisoner records.

Linking up with NCRB systems forms part of how our setup func-tions under current rules. Connection to Aadhaar was left out by design. Privacy became a core legal right after the Supreme Court ruled in the Puttaswamy case, shaping tighter limits on how Aad-haar information can be used.

## 7.2 Evidentiary Status

Still up in the air is how courts will treat images made by artifi-cial intelligence. Picture those hand-drawn suspect faces from old police shows - those are not proof, just a way to capture what some-one said they saw. Much like that, outputs from AI could serve only to guide detective work, helping sort through possible leads with-out claiming anyone did anything. Their role fits better off at trial, inside early probes instead.

## 7.3 Human Oversight

Not a single result moves forward without approval from qualified reviewers. Built right into the design, human oversight guides ev-ery step. Instead of acting on its own, the tool supports inquiry by surfacing information. Only after careful review can findings play a role in active cases. Rules stay followed because people remain in control.

## 7.4 Bias and Fairness

What we saw in age and gender patterns hints at unfair treatment. Systems already in use must face routine checks - results split by group, shared openly. Testing for slant should simply be part of rolling out any tool.

# 8. LIMITATIONS AND FUTURE WORK

This thing won't pretend it handles tasks it can't. It draws clear lines around its limits without sugarcoating. What you see is ex-actly what works - nothing hidden, nothing exaggerated. Bound-aries are set firmly on purpose. Expect no magic tricks or unseen extras beyond that frame.

Built just to show it works. Not something you can put into daily use straight away.

Just early days. The tests give a first check, yet miss what it takes to be trusted in real tasks.

Still stuck on demographics. When left unguided, it leans heavily into fairer complexions along with faces shaped like those common in the West.

When connections fail, plans can unravel. Accessing APIs pre-sumes steady internet service along with standard computing gear

- conditions absent in some places.

One step ahead could tackle how edge systems shape conditioning. Matching might improve - not by rules but learning, say with ArcFace. Combining multiple witnesses? That may need smarter merging paths. Fairer outcomes depend on more even data splits across groups. Above all else, testing live - out there with actual people doing real tasks - cannot wait.

# 9. CONCLUSION

Finding ways to test an everyday problem led us here. Could today's diffusion models actually work inside forensic processes when tools are limited? The build we made shows it can happen

- though not without limits. What matters is how those boundaries shape real use.

Testing revealed FLUX.2-klein-4b generates results much quicker than FLUX.1-dev - four seconds instead of 177 on an RTX 3060. It also uses less memory, needing only 8.4GB compared to the ear-lier version's 12GB. Because it runs so efficiently, detailed foren-sic tasks become possible using everyday computers priced near 1, 500.

Still, plenty of effort sits ahead before this method helps real cases. Early testing shows promise, though fairness across groups de-mands constant care. Laws around its use stay unclear, while trials with actual eyewitnesses or law enforcement have yet to happen. Honesty drives our talk of limits - these things weigh on every project like this. Tools needed for AI-powered sketch tech? They're out there. Trouble shows up in quieter ways: fixing skewed out-comes, testing properly, shaping laws, grappling with real-world hurdles still around the corner.

What stands out most is how we put things into practice. Running today's generative models on budget-friendly machines turned out to be fully doable - speed stayed high, making real-time interaction possible. Choice of model made a clear difference when it came to actually using them in real settings. Looking closely at laws in India revealed practical constraints worth noting. Finding hurdles ahead shaped where efforts need to go next. A live version now runs, showing what's possible when ideas take form.

Getting an idea out of the lab and into real use takes time. Field tests come first, then checks for unfair patterns, questions about laws, followed by heavy technical effort. Still - progress shows it can be done. Where police lack tools for solving crimes, especially in poorer regions, using artificial intelligence to build face sketches might just make sense. That possibility stands, despite hurdles.

# 10. ACKNOWLEDGMENTS

We thank Vivekanand Education Society's Institute of Technology for infrastructure and support. We acknowledge the developers of FLUX, Stable Diffusion, and the open-source tools that enabled this work.

# 11. REFERENCES

[1] National Crime Records Bureau, "Crime in India 2022," Min-istry of Home Affairs, Govt. of India, 2023.

[2] C. D. Frowd, P. J. B. Hancock, and D. Carson, "EvoFIT: A holistic, evolutionary facial imaging technique for creating composites," *ACM Trans. Applied Perception*, vol. 1, no. 1,pp. 19-39, 2004.

[3] S. Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "DeepFace-Drawing: Deep generation of face images from sketches," *ACM Trans. Graphics*, vol. 39, no. 4, pp. 1-16, 2020.

[4] S. Yu, J. Liu, and K. M. Lam, "Semi-Siamese network for sketch-to-photo retrieval," in *Proc. IEEE CVPR*, 2021, pp. 8007-8016.

[5] J. Wang and L. Zhang, "PI-GAN: A novel generative adversarial network for photo-realistic face image synthesis from sketches," 2022.

[6] Y. Li, X. Chen, F. Yang, et al., "DeepFacePencil: Creating face images from freehand sketches," in *Proc. ECCV*, 2020,pp. 603-620.

[7] S. Ghosh, A. Hiranandani, O. Kumar, and R. Nachane, "Do generative AI models output harm while representing non-Western cultures," arXiv:2407.14779, 2024.

[8] R. Leyva, Y. Wang, and T. Zhu, "Demographic bias effects on face image synthesis," in *Proc. IEEE CVPRW*, 2024, pp. 3892-3901.

[9] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," arXiv:2302.05543, 2023.

[10] C. Mou, X. Wang, L. Xie, et al., "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," arXiv:2302.08453, 2023.

[11] J. Zhang, Y. Liu, and H. Chen, "FluxSchell: High-fidelity multimodal generation with sparse text and visual inputs," 2024.

[12] Li, Y. Zhang, and M. Wang, "Bridging the gap between text and face images with contrastive learning," 2024.

[13] D. Chen, L. Wang, and X. Liu, "Beyond the sketch: A deep learning framework for text-guided face synthesis," 2023.

[14] K. Ka¨rkka¨inen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age," in *Proc. IEEE WACV*, 2021, pp. 1548-1558.

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE CVPR*, 2019, pp. 4401-4410.

[16] L. Song, X. Wu, and Y. Chen, "SketchFace: A large-scale dataset for human sketch-to-face synthesis," 2024.

[17] R. Sharma, A. Gupta, and V. Kumar, "PencilSketch-to-face: Challenges and approaches in criminal identification systems," 2021.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600-612, 2004.