

Street Scenes: Public Appliances for GenAI Video in Informal Settlements

Gavin Bailey

Swansea University
Swansea, United Kingdom
g.bailey@swansea.ac.uk

Dani Kalarikalayil Raju

Studio Hasi
Mumbai, India
daniel@studiohasi.com

Jennifer Pearson

Swansea University
Swansea, Wales, United Kingdom
bagelobagel@gmail.com

Simon Robinson

Swansea University
Swansea, United Kingdom
s.n.w.robinson@swansea.ac.uk

Matt Jones

Swansea University
Swansea, United Kingdom
always@acm.org

Abstract

Generative AI is rapidly diffusing worldwide, yet access remains uneven. In informal settlements, barriers of cost, literacy and connectivity can exclude residents from AI-enabled self-expression. This paper presents Street Scenes: a public appliance for walk-up interaction with generative AI video in Dharavi, Mumbai. Inspired by the “Hole in the Wall” computers and previous Dharavi speech deployments, the system lets passers-by capture phone images, add voice-, button- and dial-based prompts, and generate short videos to view and leave locally. We report on ideation workshops, two Wizard-of-Oz prototypes and a 13-day in-situ deployment across Dharavi street locations. Findings show residents appropriating AI for play, self-presentation, small business promotion and community messaging, while also raising concerns about privacy, trust and misuse. We contribute: (1) a model for public AI appliances; (2) empirical insights into community engagement with generative AI; and, (3) design lessons for accessible, equitable and community-governed AI systems.

CCS Concepts

• **Human-centered computing** → *Field studies; Interaction devices; Empirical studies in HCI.*

Keywords

Generative AI, public appliances, creativity, informal settlements, slum communities, HCI4D, ICTD, video generation, public space technologies, community media

ACM Reference Format:

Gavin Bailey, Dani Kalarikalayil Raju, Jennifer Pearson, Simon Robinson, and Matt Jones. 2026. Street Scenes: Public Appliances for GenAI Video in Informal Settlements. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3772318.3791033>



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
<https://doi.org/10.1145/3772318.3791033>

1 Introduction

There is a great deal of hope that advances in AI will bring positive effects to low-resource communities in the Global South. Already there are inspiring examples of use of the technology to tackle United Nations Sustainable Development Goals such as improved health outcomes¹ and food scarcities.² While obviously to be welcomed, there is a sparsity of studies examining how AI might be used for end-user driven and less “worthy” uses, such as self-expression and creativity. Meanwhile, in the Global North, there is much excitement and a growing user-base for AI that enables just that sort of activity. These tools—like Sora, Veo, Udio and many others—allow people to easily generate images, video and songs.³

Such sophisticated tools for content creation and self-expression, though, remain—and are likely to remain—out of reach for many millions of people who live in the Global South. Firstly, the costs of use—for example, ₹17,500 (\$200) per month for the video-generating Pro version of Sora—are prohibitive in communities such as Asia’s largest informal settlement, Dharavi. Indeed, for most households in Dharavi, where monthly incomes typically range from ₹3,000–₹15,000, even OpenAI’s cheapest offering, ChatGPT Go (₹399 per month), launched in August 2025, represents between 3–13% of household income, making it effectively unaffordable; thus, free tiers remain the only viable option, albeit with their strict limitations on capabilities and frequency of use. This underscores how commercial AI pricing structures reinforce existing digital divides in low-income urban communities. Even with a service cost of zero, in many Global South regions the cost of data packages still reduces low-resource users’ access to well-established services such as video streaming [29]. Then there are long-standing and enduring costs and access issues with regards to grid power for charging during the day [44, 46–48]. Because of such constraints, people often extend a phone’s charge by turning off non-essential features and reducing screen brightness or connectivity [22, 38]. In such a context, AI services that demand heavy on-device computation or substantial data transfer are likely to be unattractive.

A further barrier, regardless of the economics of AI, is the average levels of formal educational attainment, including textual and digital literacy, that are relatively low in places such as the

¹E.g., the PROMPTS maternal health chatbot: <https://jacarandahealth.org/prompts/>

²See: <https://www.theguardian.com/world/2024/sep/30/high-tech-high-yields-the-kenyan-farmers-deploying-ai-to-increase-productivity>

³Sora: <https://openai.com/sora/>; Veo: <https://deepmind.google/models/veo/>; Udio: <https://www.udio.com/>

informal settlements or slums we have worked with in this research. Given that the current versions of most AI tools are firmly targeted towards literate, tech-savvy users, it is unlikely an ordinary slum dweller will be able to navigate their use. Furthermore, as such people will usually only have access to a smartphone—not a desktop, tablet or laptop computer—the small-screen form factor may well exacerbate usability issues.

In 1999, Mitra’s “Hole in the Wall” experiment gave nearby slum dwellers access to a shared PC, a technology otherwise far beyond their reach [28]. Building on this idea, the 2016 UK-India StreetWise project deployed public voice-driven search appliances across Dharavi [35], offering services such as those provided by early Alexa and Google Home devices. While such consumer devices were (and remain) uncommon in Dharavi, these public deployments allowed residents to experience emerging technologies and revealed usability and value issues relevant both locally and more broadly.

Our work extends this trajectory to generative AI (GenAI) video through the design, evaluation, and deployment of *Street Scenes*, a walk-up appliance that lets people create and share short videos via images, voice and physical prompts. We trace its development from initial designs to in-situ deployment in Dharavi, highlighting three key contributions:

A public appliance model for GenAI: We go beyond prior work by demonstrating how generative AI systems can be situated in public, kiosk-like appliances rather than on personal devices. From our deployment we surface forms of collective engagement, shared discussion and appropriation that would be otherwise invisible in private use.

Empirical insights into appropriation in an informal settlement: Drawing on workshops, Wizard-of-Oz prototyping and a 13-day deployment, we detail how residents of Dharavi engaged with generative video for play, self-expression, aspirational identity, business promotion and community messaging, while also raising concerns about misuse, trust and privacy.

Design lessons for accessible and accountable AI systems: From our findings, we suggest actionable implications for designing public-facing AI: lowering barriers to entry, supporting collective and inclusive use, embedding governance and moderation, and aligning sustainability with existing local economies.

Together, these contributions underscore the role of creativity, play and enterprise, alongside “worthy” applications like health and education, in shaping how GenAI might be integrated, resisted, or reimaged in lower-resource settings.

We begin the paper with a review of related literature to identify the way our work has been shaped by those of others and the research gaps we have tried to address. Then, we began the design journey by considering drivers from two sets of workshops in Nairobi, Kenya and Mumbai, India that led to our focus on GenAI video. We follow this with a description of a Wizard-of-Oz prototype that was explored—and iterated upon—with the help of community members in India in workshop sessions. This work led to the fully-working prototype deployed in two locations in Dharavi.

2 Related work

While generative AI has been broadly studied in the Global North (e.g., [14, 15, 20]), there is comparatively limited understanding of

how such systems might be appropriated across more resource-constrained contexts in the Global South by end-users. The focus of this paper is on the potential value of generative AI systems in informal settlements – in particular, when deployed in public, rather than private, settings.

2.1 Generative AI and informal settlements

Since the start of the recent AI boom with the 2022 release of ChatGPT there has been some interest in the use of AI in more resource-constrained settings, particularly outside of the Global North. However—and importantly—the majority of this work has been produced with the aim of analysing or understanding slum infrastructures, residents and behaviours, but with little involvement of the residents themselves. For example, researchers have explored the use of deep learning for mapping buildings and terrain [39], migration and deprivation [1] and urban planning [49]. Others have reported on the failures of AI models to localise to worldwide languages, cultures and contexts – Qadri et al. [37], for example, illuminate this issue with everyday examples from across South Asia, while Jones et al. [17] focus on communities in South Africa and India, emphasising both the creativity and agency of slum dwellers, and the current narrow lens of GenAI systems and their training data [2, 30, 33].

In this paper we take a different approach, seeing slum dwellers as both users and co-researchers. Our work is therefore more aligned with that of Mim et al. [27], who explored the use and impacts of text-to-image AI tools on the digital image-making practices of Bangladeshi image practitioners and amateur artists. They found widespread use, with local adaptations such as organised account sharing, and a variety of impacts and challenges (such as English proficiency, lack of local contextualisation, distortion or misunderstanding of local culture, amongst many others). Here, we move beyond such expert users, or the SMB users described by Ankrah et al. [3] and instead deploy a GenAI service in a public setting, opening access to everyday passersby.

2.2 Public and shared technologies in low-resource contexts

Nemer’s 2022 book [31] explores how slum dwellers in Brazil’s favelas use technology in creative ways. Often this is through adaptation, selecting appropriate parts of new technologies and deploying them in imaginative ways that better fit their own context. Particularly when infrastructures are shared, or in public spaces, adaptation is often not a side-effect, but a primary mode of use. Work on charging kiosks [5] and phone-sharing [42] has demonstrated how such shared infrastructure creates opportunities for communal appropriation and creativity. Our own work embeds appropriation throughout: users remix media, prompts and other inputs to generate entirely new video content.

A well-known—and striking—example of the potential benefits of technology in public spaces was Mitra’s “hole-in-the-wall” studies [28], where making an internet-enabled computer publicly available embedded in the wall of an Indian university campus led to use by many people who would not otherwise have had access. HCI researchers have used a similar technique to explore the benefits of language AI systems in informal settlements, finding through

a Wizard-of-Oz study [40] and a larger-scale deployment [35] that public space deployments of such technologies are valuable, albeit with pointers to future adaptations required for local contexts. We adopt this approach in our own work, beginning with workshop-based studies of a simulated system and refining iteratively, ultimately leading to fully-realised public space deployments.

A key contrast between our research and that of these previous projects is the shift from information access towards creative authorship. Our work, then, links with previous efforts such as StoryBank [13], where rural Indian villagers used cameraphones to capture and compose stories that were then shared on a situated village display. Similarly, Sangeet Swara [45] allowed users to record and browse audio messages via an IVR-driven voice forum. Connecting these two themes, our own research draws on the creativity and composition seen in [13, 45] with the broad access and inclusion benefits of a public space device as pioneered by [28].

Although there have been relatively few studies on public technologies in low-resource contexts, there is more extensive literature that draws on work in the minority world. Previous research, then, provides complementary insights into how shared, situated technologies engage communities, sustain participation, and negotiate moderation and governance (e.g., [19, 25, 26]).

2.3 AI interaction in shared spaces

Turning now more broadly to previous explorations of AI systems and services in public spaces. Taking a wide perspective of both what constitutes AI and public space, one could argue that AI systems are fully embedded in public spaces. Consider, for example, self-service checkout systems, license plate detection and vehicle tracking platforms, customer service chatbots and the like, and it is hard to escape the wide reach of artificial intelligence. Here, however, we restrict to a narrower interpretation that considers only artefacts or installations that are in publicly-accessible places and that explicitly use AI in novel or imaginative ways – most prominently generative AI services. Many such prior works have had an artistic focus in controlled settings, such as museums. For example, Long et al. [21] review the use and potential for AI in public spaces, with a focus on co-creation. Their installations range from collaborative dancing to robot improvisation, music composition to storytelling, and they conclude with a broad set of design principles to make such systems more successful. Such projects also often take a deliberately performative group-based interaction approach. For example, Ronchi and Benghi [41] created an interactive environment where audience members participate in constructing a new music composition with the aid of AI. In contrast, our deployments are situated in busy streets, with an individual walk-up-and-use design.

3 Motivation

In December 2024 and February 2025, we convened a series of workshops in India and Kenya, respectively. The aim was to gather views from low-resource communities on how the future of interactive AI might best serve them. Over a total of five day-long events (three in Mumbai, India; two in Nairobi, Kenya), we worked with 60 participants who lived and worked in informal urban and rural settlements. Almost no one in the groups had experience or had heard of AI, giving us access to refreshing perspectives on the technology [36].

One of the workshop activities involved asking groups of participants to envisage a “magic thing” [16], an AI device or service which could do anything they imagined for whatever purpose they desired. To visualise these inventions, we used various forms of video GenAI. That is, after a round-table discussion, the group described the AI service and then we converted this description into an AI prompt to visualise the “magic thing” in action. The video was then used as a basis for further discussion; the limitations and the successful elements of the visualisations helping participants refine and innovate on their ideas.

While the ideas created were fascinating, participants were also keen to explore how GenAI could be used in itself. For example, one person in the Nairobi sessions was an egg-seller who asked us to use the tool to create video adverts to promote the sales of her stock (see Fig. 1); and, another who earned her living from providing fashion wigs was excited to see the tool generate potential new designs. Other participants were keen to visualise how changes to their community environment and behaviours could improve daily lives. Others wanted to use the GenAI to inspire them in terms of new forms of dance and entertainment.

So, while we went into the workshops with a view to using video GenAI tools as one of the methods for eliciting ideas for future AI services, we left them seeing that providing access to these tools might in itself provide “magic things” for users.

4 Wizard-of-Oz prototyping

4.1 Prototype 1: initial design

With the inspiration from the Dharavi and Nairobi workshops, allied to a detailed review of previous public space technologies in slum settings, we developed a Wizard-of-Oz prototype as shown in Fig. 2. To generate a video, a user first adds up to three images from their own phone, either via Bluetooth or by holding their phone in front of the device’s camera, which takes a photo of the mobile phone screen. These images are used as a basis for a prompt to a video generation AI service, following the process described below. In addition, surrounding the screen are a series of push-button and slider-based controls to provide additional dimensions that are embedded into the prompt. So, a user can press the buttons to prompt for a video that has one or more of these features: fun; an advert; of value to the wider community; and, a surprise. Similarly, by using the sliders the user can request a video that is: more or less realistic or cartoon in nature; based in the future or more in the past; and, that is more or less panoramic or close-up in shot composition. We chose these prompt features from a combination of the requests we saw in the prior workshops in Nairobi and Mumbai and a review of styles seen in the interfaces of GenAI video tools such as OpenAI’s Sora. In addition the prototype also enables a user to record a spoken narration of what they would like to see in the video.

For this first prototype, we leveraged commercially available state-of-the-art generative AI tools from OpenAI, deploying them through a Wizard-of-Oz approach⁴. In practice, this meant that while the system appeared fully automated to participants, a human operator executed each step of the pipeline behind the scenes. Guided by findings on issues in the use of Large Language Models

⁴Services used: ChatGPT: <https://chatgpt.com/>; Sora: <https://openai.com/sora/>; Text to Speech: <https://platform.openai.com/docs/guides/text-to-speech>



Figure 1: Top: a still image from a Sora-generated video in response to a Nairobi participant’s request for a video to advertise her eggs (for later posting on TikTok). Bottom: output from Dharavi participant’s request for an advertising video during our Wizard-of-Oz studies of Prototype 1 (see Section 4.1), with the prompt “The name of my shop is Narkhar Parsan Mart, and I need to advertise it, so you can get a video of that type.” and a photo of the participant’s family standing next to a yellow wall.

(LLMs) in HCI research [34], we provide all prompts and scripts used in Appendix A and in the supplementary material for this paper. The pipeline for generating a video from user input proceeded as follows:

Speech recognition: User speech (if present) was transcribed in the original spoken language using Google Cloud Speech-to-Text.⁵ If the original language was not English, the text was translated.

Script generation: The transcribed text, together with user-provided images and interface inputs (i.e., slider and button positions), was passed to ChatGPT (model “GPT-4o”). We prompted ChatGPT to produce a detailed script suitable for generating a Sora video. The prompt design was itself

⁵See: <https://cloud.google.com/speech-to-text>

Table 1: Initial Wizard-of-Oz prototype configurations indicating the inputs a participant could provide to prompt a generative AI video. Each participant in the session used one configuration (A–E). Configuration E was the same as that in C except that after seeing the generated video, a participant was able to refine their voice prompt to improve the output.

| | Voice | 1–3 images | Buttons / sliders | Iteration |
|---|-------|------------|-------------------|-----------|
| A | ✓ | - | - | - |
| B | ✓ | ✓ | - | - |
| C | ✓ | ✓ | ✓ | - |
| D | - | ✓ | ✓ | - |
| E | ✓ | ✓ | ✓ | ✓ |

optimised by asking ChatGPT to propose effective prompt structures for Sora.

Video creation: The generated script was submitted to OpenAI’s Sora to produce a 15-second video. Since at the time of our study Sora permitted only a single image input, the Wizard selected the image with the highest number of people depicted (or the first image if none contained people) and attached it to the prompt. Any images of children were not chosen to be added to the prompt.

Audio narration OpenAI text-to-speech (voice=“shimmer”, model=“tts-1”) processed the same script to synthesise a spoken narration of the story.

Assembly and playback: Once Sora returned the generated video, the Wizard downloaded it, synchronised it with the narration and wirelessly transferred it to the prototype. The final output—a video with accompanying audio—was presented to participants on the prototype’s screen.

We choose to use a Wizard-of-Oz approach and to assemble the system using rapid-prototyping components and a series of external web services in order to be able to quickly create a prototype that could be shown to users—and potentially radically changed—without spending a great deal of time and resources before this validation had occurred [7]. This approach allowed us to investigate how participants imagined and interacted with an emergent AI technology before a fully functional version was feasible, guided by useful recent findings on interactive AI prototyping [43], and successfully used by others in Global South contexts [40].

4.2 Evaluating the initial design: method

All of the studies reported in this paper were approved by our institution’s ethical review board. The final study designs were created in consultation with a community representative.

We recruited 15 Dharavi residents (9 women, 6 men), aged 19 to 58 years ($M=36.6$), to explore the use and value of the initial prototype. Participants attended in three groups of five people, with each session lasting approximately 90 minutes. After explaining the purpose of the study and obtaining informed consent, we spent approximately 15 minutes individually with each participant following the procedure below.

Background and expectations: We collected demographic information and prior experience with AI and social media.

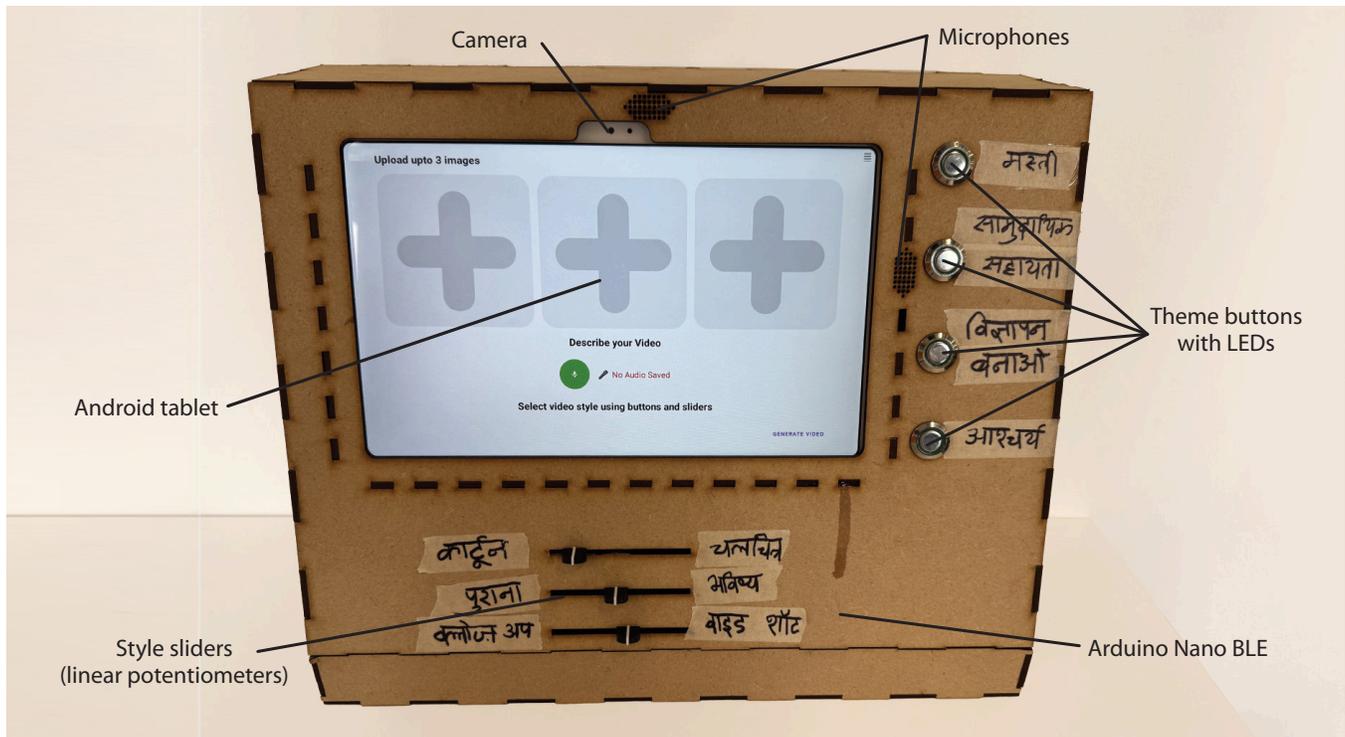


Figure 2: The first Street Scenes prototype. Users add 1–3 images from their phone (via Bluetooth or the prototype’s camera), and optionally use the physical controls to indicate the desired themes (buttons) or styles (sliders) of the video. Translations of the Hindi theme button labels, from top to bottom – Fun; Community help; Advertise; Surprise me. Translations of style sliders, from top to bottom – Cartoon ↔ Photorealistic; Old fashioned ↔ Futuristic; Close-up ↔ Wide shots. An optional spoken narration can be recorded to describe what the user would like to see in the video. The generated video is then shown on-screen when ready. The prototype software is an Android app running on a Samsung tablet.

Participants were introduced to the concept of a GenAI appliance and its public setting. *Before* seeing and using the prototype, they were asked to imagine potential uses of such a device for themselves and their community.

Prototype interaction: Each participant was assigned one of five system configurations, with various prototype features enabled or disabled (see Table 1). In total, over the three sessions and 15 participants, three participants tested each configuration. For image upload, participants were asked to do so via Bluetooth transfer from their device as a first choice.

During and after video generation: Video generation took around two minutes on average. While waiting, participants were asked what they expected the system to produce and whether they could suggest alternative controls or refinements for the interface. After watching their generated video, they rated it (1–10, 10=high), described how they would improve it, and reflected on whether they would be happy leaving the video on the device for others to view. As a proxy for personal value, we asked whether they would transfer the video to their phone and, if so, whether they would share it further.

Post-task: After completing the task, each participant waited in a separate room to avoid influencing others. Once all five participants had completed the system interaction, we conducted a short group discussion, thanked them for their time, and provided a cash honorarium (₹500; ≈\$6).

Our study was designed as an exploratory, formative evaluation. The group of 15 participants was chosen to provide a diversity of perspectives to identify use patterns and common issues. Conducting the study in small groups facilitated community participation, while individual sessions ensured that each participant could interact with the prototype independently. The goal was not statistical generalisation or controlled comparison of interface versions, but rather to surface expectations, interaction breakdowns and design opportunities to guide subsequent iterations of the system.

Three of the research team took notes of participants’ answers and of their observations of the ways the participants interacted with the prototype. These notes were used in group analysis sessions to classify suggested uses and issues, additional and alternative features, as well as summarise quantitative data (such as the rating of the video).

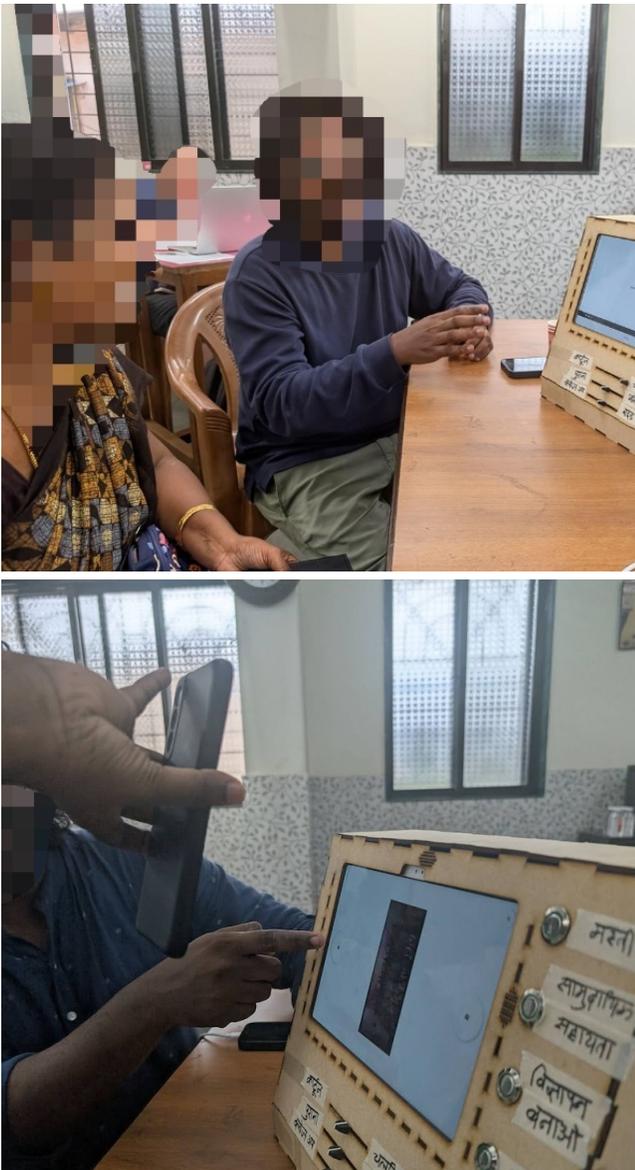


Figure 3: Top: study setting – a local community hall in Dhara-avi. Participants were situated in front of the prototype; two researchers sit to their right (behind the camera here) and left, and the Wizard sits separately. Bottom: the second Street Scenes prototype. The option to upload images via Bluetooth has been removed, and images are instead captured directly by the device’s camera.

4.3 Findings from the initial design

4.3.1 Context and technology. While all participants could speak and understand Hindi and Marathi, only eight were fully literate, and three identified as semi-literate. Fourteen participants owned Android smartphones, with one owning a featurephone; in two cases, phones were shared among family members. Five participants’ phones had damaged screens, and for two-thirds, a phone

was the only computing device in their household. While familiar with WhatsApp and video consumption, awareness and direct experience with generative AI were limited, though most had encountered biometric systems (for example, via India’s national identity system, Aadhaar).

4.3.2 Imagined uses of video-based generative AI. Before seeing or interacting with our prototype, participants expressed enthusiasm about potential applications. Suggested uses included: creating recipe videos from food photos; teaching children through religious or cultural stories; advertising small businesses; animating photos of deceased loved ones; and, generating greeting-card style videos for sharing.

Eight participants imagined videos primarily as tools for informing or inspiring others (e.g., advertising, teaching). Five saw value for personal use (e.g., entertainment, inspiration). Two people described particularly interesting use-cases. One wanted to see a video based on his photo showing more of the surroundings (“*behind and out-of-sight of the camera*”). Their motivation was that their phone had limited storage and they were not keen to make their own videos in situ. Another spoke of wanting to see themselves in places that they cannot travel to because of limited resources.

While most participants spoke of time-independent uses, one of the participants suggested that the use of the tool would vary depending on the time of day: “*In the morning people would like religious video to inspire them. Then after a day at work they might have anger about another and upload an image of them and ask the system to punish that person. In the night, some might use it for pornography*”.

4.3.3 Use of images, buttons, sliders and voice prompting. The twelve participants using image-enabled configurations (B–E in Table 1) uploaded, on average, 1.75 photos. In configurations where theme-selection buttons were available, participants chose exactly one theme each, with no theme emerging as dominant. Slider logs showed preferences leaning toward photo-realism, futuristic styles and wide shots. Nine of the 12 participants who used a configuration that enabled voice prompting (A–C, E) chose to do so.

Three participants proposed additional controls: a slider to classify video appropriateness (e.g., Unrestricted through to Adult rating); a button to blend related YouTube content into their content; and, a way to highlight focal points in uploaded images for the story generator.

4.3.4 Quality and value of the generated videos. The average video rating was 5.3/10 (s.d. 3.5), with no systematic differences by prototype configuration. One group showed extreme scores (four ‘1’s, one ‘10’), mostly due to errors in speech-to-text transcription or translation that produced irrelevant prompts (e.g., “floods” became a cat drinking tea). Excluding these, the mean rose to 6.4 (s.d. 3.1). In general, speech input proved unreliable: only 3 of 9 audio prompts were fully accurately transcribed, with background noise (such as ceiling fans and heavy rain outside) often rendering audio unintelligible.

Image-based configurations (B–E) posed other challenges. Participants struggled to select coherent photo sets, and Bluetooth transfer was slow (~1 minute per image), requiring assistance in most cases. Videos often diverged from uploaded images, as Sora quickly morphed them into synthetic content. Half of image-users wished their own photos were animated more directly.

On sharing, nine of 15 participants were comfortable leaving videos on the device for public viewing; six declined due to privacy or relevance concerns. Nearly all wanted to take videos away for sharing, typically via WhatsApp.

4.3.5 Summary. Overall, participants showed enthusiasm for video-based generative AI, suggesting imaginative personal and community applications. At the same time, the study revealed significant constraints: low textual literacy, infrastructural limitations (e.g., damaged phones and shared devices, issues related to Bluetooth), and inadequate AI services (e.g., in speech-to-text and image fidelity) which shaped both what participants could do and how they valued outcomes. These findings highlight the tension between the promise of generative AI and the realities of situated use in lower-resource contexts, suggesting design opportunities for more resilient, accessible and locally relevant systems.

4.4 Prototype 2: refining the design

After the first workshop, our team met to reflect on the findings and rapidly re-prototype the system. Several key limitations identified in Prototype 1 guided our design revisions: (i) participants struggled to select coherent sets of images; (ii) many expressed a desire for their photos to be animated more directly; and, (iii) Bluetooth transfers proved slow and difficult.

Based on these insights, we introduced two distinct video production modes in the second prototype:

Inspire me: Users were informed that the 1–3 images they provided would not appear directly in the video, but instead would serve as inspiration for an AI “film-maker.” That is, we told participants that the system would analyse every image in high detail but not attempt to use the actual visual image in the generated videos. This option was designed to reduce the burden on participants to construct a full narrative from their photos while, pragmatically, also accommodating the technical constraint that Sora could only process one image input.

Bring to life: Users uploaded a single image, which the system attempted to animate more fully than in Prototype 1. This directly addressed participants’ requests for their own photos to be animated, rather than merely used as prompts.

To support these two options, we iterated on several prompt structures. One of the main changes was to add, for the Inspire me mode, a new step of generating text-based image descriptions rather than using the images themselves in the final video generation step. As with the first prototype, we used ChatGPT to co-design the prompts, asking it to suggest effective formulations tailored to the two use cases. The final prompts and video generation processes are detailed in [Appendix A.2](#).

Finally, to overcome the difficulties of Bluetooth image transfer observed in the first study, we changed the system solely to capture images directly using its built-in camera. While previously we had viewed this as a fallback method, in practice it proved far more reliable and accessible. Participants could simply hold their phone up to share the image (see [Fig. 3](#)), eliminating delays and technical barriers.

4.5 Evaluating the second design: method

The refined prototype was used by a further 15 participants (5 men, 10 women) between the ages of 18 and 60 ($M=27.5$) to test the iterated design. Nine of the participants were asked to use the “inspire me” version and six the “bring to life” option. We asked more people to use the first configuration than the second as we felt this system had the potential to afford richer video composition insights. For both configurations, a participant could also choose to use none, one or more of the button prompts and sliders and / or record an audio prompt. The interactions and participant responses were recorded and analysed as in the prior study, above.

4.6 Findings from the second design

4.6.1 Context and technology. All participants were again residents of Dharavi. Technology access, literacies and competencies broadly mirrored the previous cohort (none had participated in that study).

4.6.2 Imagined uses of video-based generative AI. Before interacting with the prototype, participants suggested applications similar to those voiced in the first study, including: ideation and inspiration (e.g., generating new mehndi (henna tattoo) patterns); community communication (e.g., announcing celebrations or highlighting local problems); advertising; and, pornography.

4.6.3 Use of system features. Participants found the two top-level options—“inspire me” and “bring to life”—easier to understand than the single multi-configuration prototype tested previously. This was evident in how they identified images more readily and described them as sets of content, rather than struggling to piece media into coherent stories. The revised image upload method proved substantially faster and smoother than Bluetooth transfer, directly addressing frustrations from Prototype 1.

As in the earlier study, participants generally selected only one theme button, with ‘Fun’ being the most frequently chosen. Slider use indicated preferences leaning toward cartoon styles, futuristic presentations and wide shots. Although audio prompting was available, only two of the 15 participants opted to use it. In asking participants why they did not want to use this feature we were given three main reasons: difficulties a participant had in framing an utterance as a prompt; shyness or embarrassment in having to speak to the “box”; and, ambient background noise.

Seven participants suggested additional control ideas, including: buttons / sliders for emotional tone (e.g., happy, sad, tired); and, buttons for motivational content (e.g., inspire me, motivate me, give me new ideas). These suggestions point toward a desire for more expressive, affective control over video generation beyond simply aesthetic and structural dimensions.

4.6.4 Quality and value of the generated videos. The average video rating was 6.8 out of 10 (s.d. 2.3). Participants’ most common requests to improve the videos were: faster video generation times; and, greater fidelity when animating a personal photo, as Sora typically maintained the original image for around five seconds before diverging into less-related content. Sharing practices remained consistent with the first study: 11 of 15 participants said they would leave their videos on the appliance for public viewing, and all reported they would take the content to their phones for private sharing (most commonly via WhatsApp).

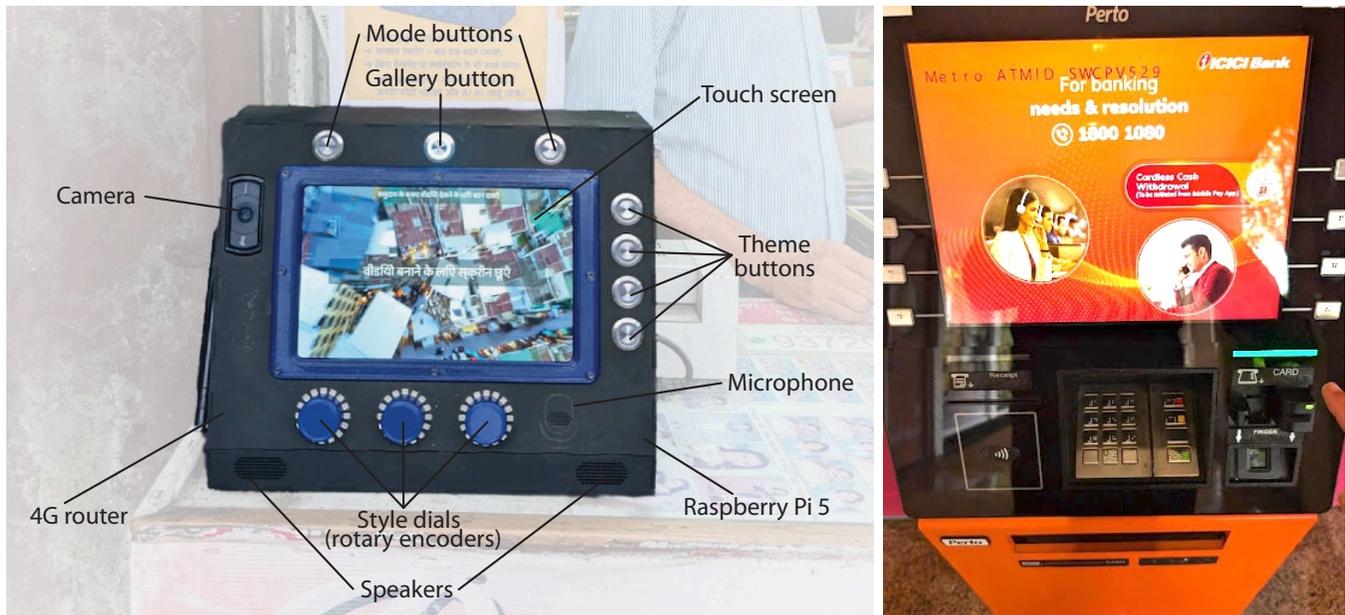


Figure 4: Left: One of the two Street Scenes appliances deployed in Dharavi. Figures 5 and 6 show user interaction steps and the prototype in the deployed locations. Right: A typical ATM in Mumbai. This type of combination of physical elements combined with touch-screen interactions along with the use of LEDs to guide interaction inspired our final prototype design.

4.7 Discussion

Our two rounds of Wizard-of-Oz prototyping revealed both opportunities and challenges in introducing public generative AI technologies into lower-resource urban contexts. Rather than reviewing results in detail again, here we draw out key lessons that helped shape the design of the deployable prototype that we describe in the next section.

4.7.1 Expectations and imaginative uses. Participants consistently envisioned the appliance as more than a novelty, imagining uses spanning personal play, aspirational identity, small business promotion, and community messaging. Importantly, they located value not only in outputs for themselves but in the potential of the device as a shared street-level medium.

4.7.2 Interaction design trade-offs. Prototype 1 highlighted the difficulty of supporting multi-modal prompting in resource-constrained settings: Bluetooth transfers were slow, speech recognition often failed in the noisy environment, and participants struggled to compose coherent narratives from multiple images. Prototype 2 simplified this through the “bring to life” and “inspire me” modes, and removed Bluetooth to rely solely on direct camera capture. These changes reduced complexity and improved usability, highlighting the importance in this context of robust and legible interactions over high-degrees of flexibility.

4.7.3 Tensions around fidelity and expression. Across both studies, participants requested closer alignment between their own images and the generated outputs. At the same time, some sought richer expressive controls for emotion, style and motivation. This points

to a central design challenge: balancing fidelity to user-provided media with the creative possibilities platformed by generative AI.

4.7.4 From individual to public value. Most participants expressed willingness to share videos publicly, framing outputs as valuable not only for themselves but for others who might use the service. But concerns about privacy and potential misuse emerged as well. This highlights the need to support hybrid public / private practices, where users can choose how outputs circulate.

4.7.5 Grounding AI development. Together, the two prototypes illustrate how grounding generative AI design in local practices and constraints can lead to gains in usability, while also reminding us of the potential for deeper design opportunities beyond purely technical fixes. Much research has shown the need for culturally and linguistically diverse data sets to train LLMs [17], but these workshops reiterate the need to go further: to build, evaluate and test systems within such diverse contexts.

5 From Wizard-of-Oz to deployable prototype

Following the two workshop studies, we returned to the lab to construct a fully-functioning Street Scenes prototype suitable for in-situ deployment in Dharavi. The goal was to translate the Wizard-of-Oz experience into an independent, street-deployable system that reflected design lessons learned from Prototypes 1 and 2.

5.1 Hardware and physical design

The deployed Street Scenes prototype, shown in Fig. 4, was based on the refined design described in Section 4.4. The design’s components and their purpose are described in Table 2. The final design of the interface drew inspiration from other kiosk-like machines

commonly seen in Dharavi, such as ATMs (see Fig. 4, right). Similar to these, the prototype combined digital and physical controls, and incorporated illuminated elements to guide user attention.

Several refinements directly responded to workshop findings. To improve accessibility, spoken instructions and confirmations of actions in Hindi supported participants with limited literacy [24]. Residents of Dharavi, who represent diverse cultural, religious and regional backgrounds, communicate in a wide range of languages, including Hindi, Marathi, Bengali, Gujarati, Kannada, Urdu and Tamil. While many residents use their native languages in interactions with fellow speakers, Hindi functions as the lingua franca across communities.

Interface legibility and use was also improved by dials replacing sliders, making controls more uniform and ensuring each new user began with neutral settings (avoiding leftover slider positions from previous interactions). Friction in the interaction flow was minimised by using the camera-based image capture of Prototype 2 rather than the slow and error-prone Bluetooth transfer of Prototype 1. When choosing and integrating the appliance’s microphones, we were mindful of the need to reduce the impact of ambient street noise. We did not implement a direct way for a user to “take-away” their generated video – this reduced the risk of any problematic videos being easily shared. In that regard, for content moderation, we tasked device caretakers with reporting video issues to us to action (see Section 6.1 for details).

5.2 Software and interaction flow

The Street Scenes system software provided a complete end-to-end interaction pipeline, as described below. The software was written as a Python application running on a Raspberry Pi 5. Figure 5 illustrates each stage of the interaction process, from initial welcome and consent to video viewing and rating.

Introduction and consent: On first interaction, the system explained its purpose as a research project and asked the user to actively consent or decline participation (see Fig. 5 (a–b)). This explanation (in Hindi) was both spoken and shown in writing.

Video generation options: Users chose between two top-level modes carried over from Prototype 2 (Fig. 5 (c)):

Inspire me (IME): Upload 1–3 images to inspire a story. The images are not directly used in the video but are used as inspiration in the same way as described in Prototype 2.

Bring to life (B2L): Upload a single image to be animated. After adding images, users were prompted to optionally select a theme button, adjust styling dials and / or record an audio prompt (Fig. 5 (d–i)).

Prompt construction: Inputs (image(s), optional audio description, button and dial selections) were combined into structured GenAI prompts. The prompt construction was done by populating templates via Python code (see Appendix B). These were submitted to Veo⁶ (model=“veo-3.0-generate-preview”), which at the time of deployment offered

Table 2: Design components and corresponding functionality of the deployed Street Scenes prototype.

| Component | Function |
|------------------|---|
| Raspberry Pi | Drives the display and interaction components. Uses the adjacent 4G router for connectivity (via SIM or Wi-Fi as available) |
| Physical buttons | Select generated video types (Idea, Community Support, Advertisement, Fun) |
| Rotary dials | Adjust stylistic dimensions of generated videos (e.g., realism ↔ cartoon, past ↔ future, panoramic ↔ close-up) |
| Camera | Positioned to the side of the box to allow capturing images from a user’s phone without obscuring the display |
| Microphone | Records users’ spoken descriptions of the desired video |
| Speaker | Plays back the generated video’s audio and provides system-generated spoken instructions |
| Public gallery | Allows viewing videos left by previous users |

more stable integration than our previous choice of video GenAI service (Sora). Veo generated 8-second videos (the maximum permitted length via its API). Full details of the prompt structures used are given in Appendix B.

Safety check: We used Google’s SafeSearch API and the DeepFace facial recognition framework to analyse all uploaded images. Any flagged content (e.g., nudity, children) resulted in an error message, and a prompt to retry. Although in our view the public space placement of the box made such inputs unlikely, this safeguard provided additional assurance.

Output and evaluation: After video generation, the system played the result, and asked users to rate the video (1–5 stars, 5=high). Participants could then choose to save their video to the public gallery, making it available for others to view on the device (Fig. 5 (j–k)).

6 In-situ deployment

6.1 Method

To move beyond the earlier controlled workshops, we conducted an in-situ deployment of the working prototype in Dharavi. Our goal was to observe how people encountered and used the system in everyday public settings, to understand the kinds of videos they produced, and to explore their reactions to both the process and the outputs.

With the help of a local community member, we identified two deployment sites that offered contrasting yet complementary social contexts. One prototype was placed in a small photocopying shop, where customers often had time to spare while waiting for their jobs to be completed. The shop also offered stable Wi-Fi, which made it a reliable location for longer-term deployment. The second prototype was located in a sugarcane juice kiosk situated opposite a church and community centre – an area with particularly high foot

⁶Services used: Veo: <https://deepmind.google/models/veo/>; SafeSearch: <https://cloud.google.com/vision/docs/detecting-safe-search>; DeepFace: <https://github.com/serengil/deepface>

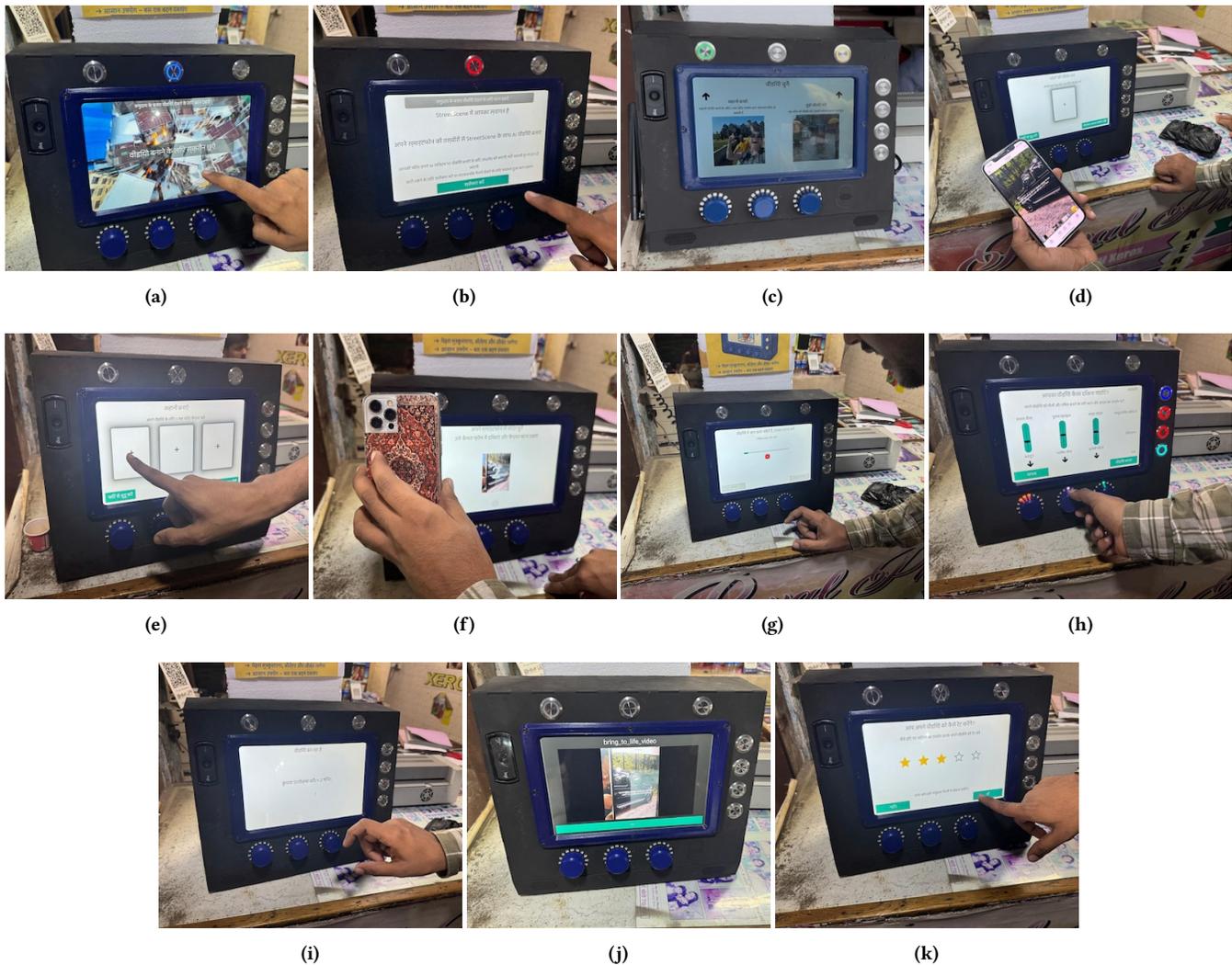


Figure 5: Street Scenes interaction sequence. From top left to bottom right: (a) Welcome screen; (b) Consent screen; (c) Service selection screen; (d) Bring to life mode selected, with the user choosing an image to upload (on their phone screen); (e) Inspire me mode selected; (f) User holds their phone in front of the device’s camera to upload an image. Users can optionally choose to: (g) record an audio prompt; and / or, (h) select a theme and adjust the style dials. The actions of the buttons and dials are indicated via on-screen labels. Once the user has confirmed their inputs, they see: (i) A holding screen requesting the user to wait 1–2 minutes for their video to be created; (j) The video playback screen; and, lastly, (k) A rating screen, with an option to allow the video to be stored on the device’s public gallery.

traffic. This site was chosen both for the density of passers-by and because customers typically paused to drink and converse, creating opportunities for engagement. Both shops were of the predominant type seen in Dharavi, having counters that are open to the street, rather than buildings that a customer enters (see Fig. 6). In both cases, we explained the study purpose to the shopkeepers, who agreed to host the prototypes and to help orient customers to the system. Each shopkeeper received ₹4,000 (approximately \$45) as compensation for hosting the appliances.

Before beginning the deployment, we carried out pilot trials in the two shops (roughly 1–2 hours at each location), experimenting with positioning the appliances to maximise visibility while avoiding obstruction of daily business. We also identified and implemented necessary minor adjustments, such as adjusting text sizes for legibility and raising the volume of audio prompts to overcome street noise.

The photocopy shop prototype was deployed for thirteen days and, on average, nine hours a day. The juice kiosk prototype was deployed for five days, with a shorter daily operational window of on average five hours, reflecting the constraints of unreliable



Figure 6: The Street Scenes deployed prototypes in situ. Left: a photocopy shop; centre: a sugarcane juice shop. Right: a close-up view of the prototype as deployed in the photocopy shop.

mobile data connectivity (the shop did not have Wi-Fi). The longer deployment in the photocopy shop was made possible by the shop’s Wi-Fi connection, which was more reliable than the mobile signal available to the sugarcane juice kiosk.

During the deployment, when a user touched the appliance’s screen, the system presented a spoken and written description of the system and the study and requested consent before use. When present, shopkeepers and members of the research team reinforced this by verbally explaining the project and answering any questions. Customers and passers-by who chose to interact with the prototype were invited to reflect briefly on their experience immediately afterwards. Interviews were also conducted with the shopkeepers throughout the study to capture their observations of how people engaged with the system.

To complement these user reports, members of the research team observed deployments unobtrusively from nearby, noting patterns of interaction and contextual details such as crowding, noise, or interruptions. Throughout the 13 days, the research team maintained an on-site presence for 41 hours (out of 143 total hours of operation), balancing the need to support participants with the desire to observe naturalistic use.

The two deployed prototypes automatically logged all interactions, generating quantitative data on the number of videos produced, ratings assigned, videos browsed, etc. The system recorded a detailed log of each video request, including button presses, dial adjustments, audio recordings and captured images. These logs also included generated image descriptions, audio transcripts and the resulting videos. Generated videos were saved for later analysis, but were only available for members of the public to view when participants explicitly consented to leave them in the public gallery.

6.2 Analysis

To analyse the image, description and video content in the prototypes’ logs, we adopted an inductive approach beginning with open coding, consistent with other AI and Global South studies (e.g., [3, 18]). The primary aim was to: (i) characterise the types of images uploaded by participants by reviewing the system’s generated descriptions; and, (ii) classify the likely purposes for which

videos were being made (e.g., for communication, to advertise a product / service, etc.).

A single researcher undertook the initial coding to ensure consistency of interpretation across the dataset. Rather than pursuing inter-rater reliability as a measure of rigour, which has been critiqued as inappropriate or unnecessary for interpretive qualitative work [10, 23], we emphasised transparency, reflexivity and collaborative sense-making [6]. Following coding, the categories and interpretations were iteratively reviewed with other team members, enabling discussion, challenge and refinement until shared understandings were reached. We also triangulated across multiple data sources, including interaction logs, observational notes and shopkeeper interviews, strengthening the robustness of our interpretations (cf. [35]).

For the image upload classification, we interrogated the detailed descriptions provided by the system; on average these descriptions were 500 words long. Meanwhile, in classifying the likely purposes of the generated video, we combined our reading of the image upload descriptions, any audio prompt transcriptions associated with the request (e.g., “advertisement for garments”), details in the script provided to Veo including instructions for the audio narration, and by viewing the output video itself.

Alongside the log and content analysis, we also drew on observational notes, user reflections and shopkeeper discussions. These materials were reviewed and discussed collectively during and after the 13-day deployment, enabling the team to identify insights and practices that shaped how the system was taken up in these public settings.

6.3 Findings

We present the findings of the quantitative and qualitative analyses of the appliances’ logged data; and, then turn to the insights from observations and interviews with customers and shopkeepers.

6.3.1 Quantitative analysis of logged data. Engagement and attrition: Across the 13-day deployment, 779 sessions were initiated (i.e., when a person performed their first interaction with the appliance). Of these, 252 (32.4%) exited at the consent screen, 34 (4.4%)

Table 3: Categorisation of uploaded images, giving descriptions and representative examples in each case. Each uploaded image was classified into one of these ten categories. The distribution across categories and prototype modes is shown in Table 4.

| Category | Description | Illustrative example |
|--------------------------------|--|--|
| Individual portraits | Selfies, headshots or portraits of one person; the most dominant type. | Selfie of a young person in a plaid shirt, lens flare across their face. |
| Leisure and lifestyle | Everyday leisure, style, consumption and aspirational lifestyle. | Young person on a motorcycle at dusk holding a drink can. |
| Group / social scenes | Images of two or more people in social or community settings. | Two people standing at a market clothing stall. |
| Animals / pets | Images featuring cats, dogs or other animals. | Portrait of a domestic cat under a canopy. |
| Work / identity contexts | Images showing uniforms, professions or explicit social identity. | Man in uniform with epaulettes. |
| Food and objects | Focus on meals, cooking, trophies or personal items. | Person eating with plates of food on a charpai. |
| Romantic / family | Romantic themes, family members or elder relatives | Blurry old photograph of an elder. |
| Abstract / textual | Logos, slogans or text graphics. | Graphic reading “Passion is your true soul”. |
| Built environment and vehicles | Cars, scooters or streetscapes as focal objects. | Mercedes-Benz G-Wagon with a forest backdrop. |
| Landscapes and monuments | Famous monuments or natural scenery. | Taj Mahal under purple haze. |

at the scenario selection stage, and 63 (8.1%) during media capture without attempting video generation. 332 sessions involved browsing the public gallery (42.6%). Ultimately, 194 participants (24.9%) successfully generated a video, with 173 (22.2%) going on to rate the outputs. The “Bring to life” (B2L) scenario accounted for most completions (134 videos, 69.1%), while “Inspire me” (IME) produced fewer (60, 30.9%). Despite this imbalance, the mean video ratings were near identical (IME: 3.42 out of 5, B2L: 3.44 out of 5), suggesting comparable levels of satisfaction.

Session duration: Session lengths further reflected users’ scenario choices. Including generation time, IME sessions averaged 270s, compared to 230s for B2L. When excluding generation, IME still demanded more user interaction (152s vs. 125s) due to the option to select and submit multiple images.

Themes and stylistic controls: Analysis of theme button use shows that most participants did not select a theme (132 out of 194 sessions, 68%). Where they did, “Fun” was the most frequent (29), followed by “Community support” (13) and “Advertisement” (12). “Idea” was rarely used (8).

Across the 194 completed sessions, 111 participants adjusted at least one style, with the most common axis being photorealistic to cartoon (106 uses), followed by old-fashioned to futuristic (75) and wide-shot to close-up (58). Combinations of multiple dials were frequent (52 sessions adjusted all three styles).

Voice prompts: Voice input was recorded in 151 sessions (78% of all completions). Prompts were typically short, averaging 8.4 words, though ranging from single-word utterances to longer sentences (up to 50 words). IME and B2L participants produced similar length prompts. These data reinforce observations that voice served primarily as a lightweight supplement to images, rather than lengthy narrative scripting.

Table 4: Categorisation of uploaded images, showing the distribution across categories and interaction modes. Definitions and illustrative examples of each category are shown in Table 3.

| Category | Total % (N=266) | IME % (N=134) | B2L % (N=132) |
|--------------------------------|--------------------|------------------|------------------|
| Individual portraits | 67.4 | 58.0 | 75.3 |
| Leisure and lifestyle | 11.8 | 13.6 | 10.3 |
| Group / social scenes | 5.6 | 7.4 | 4.1 |
| Animals / pets | 5.1 | 11.1 | 0.0 |
| Work / identity contexts | 3.4 | 1.2 | 5.2 |
| Food and objects | 1.7 | 2.5 | 1.0 |
| Romantic / family | 1.1 | 0.0 | 2.1 |
| Abstract / textual | 2.2 | 2.5 | 2.1 |
| Built environment and vehicles | 1.1 | 2.5 | 0.0 |
| Landscapes and monuments | 0.6 | 1.2 | 0.0 |

Gallery engagement: The public video gallery proved an important site of secondary interaction. A total of 332 sessions browsed the gallery, spending an average of 59s exploring. Of these, 164 (49.4%) subsequently created a video, showing how browsing had the potential to seed further participation. By the end of the deployment, 151 videos had been made public (102 B2L, 49 IME), suggesting both willingness to share and the gallery’s role in sustaining engagement

6.3.2 Qualitative analysis of the appliances’ logs. Uploaded images: There were 266 images uploaded to the appliances during the deployment (132 IME, 134 B2L). From our analysis, ten categories emerged: Individual portraits; Leisure and lifestyle; Group / social



Figure 7: Example frames from generated videos during our deployment in Dharavi. Left (IME): the user uploaded a close-up image of a bike, with the spoken prompt “his royal ride, the beginning of every new story.” Centre (B2L): the user provided a selfie photo and asked to make the image dance. Right (IME): the user showed an image of a cat and asked for a “story of a cat and dog.” In all cases, the user rated the video as 5 out of 5 and chose to place it in the public gallery.

Table 5: Distribution of video purpose categories across all videos (Overall), Inspire me (IME), and Bring to life (B2L). Category definitions and example inputs / outputs are given in Table 6.

| Category | Total % (N=194) | IME % (N=60) | B2L % (N=134) |
|-------------------------------|--------------------|-----------------|------------------|
| Unclear / exploratory use | 23.2 | 20.0 | 24.6 |
| Explanation of self / objects | 17.5 | 11.7 | 20.1 |
| Creative re-contextualisation | 16.5 | 21.7 | 14.2 |
| Positive self-presentation | 11.3 | 21.7 | 6.7 |
| Advertising / promotion | 10.8 | 18.3 | 7.5 |
| Imagining a better self | 3.6 | 10.0 | 0.7 |
| Reassurance | 3.1 | 1.7 | 3.7 |
| Learning | 2.6 | 0.0 | 3.7 |
| Romantic themes | 2.6 | 1.7 | 3.0 |
| Celebration / greeting | 2.1 | 1.7 | 2.2 |
| Community assistance | 2.1 | 1.7 | 2.2 |
| Visualising places | 1.5 | 0.0 | 2.2 |
| Seeking advice / counselling | 1.0 | 3.3 | 0.0 |
| Family life | 1.0 | 0.0 | 1.5 |
| Futuristic | 0.5 | 1.7 | 0.0 |

scenes; Animals / pets; Work / Identity contexts; Food and objects; Romantic / family; Abstract / textual images; Built environment and vehicles; and Landscapes and monuments. Coding prioritised the most salient subject or theme of each description (e.g., an image of a man holding a drink while seated on a motorcycle was coded as Leisure and lifestyle, rather than Objects). Category definitions and illustrative examples are shown in Table 3, with the distribution across categories shown in Table 4.

Video purposes: Our analysis of generated videos revealed a diverse set of purposes. Some videos were created to test what the technology could do, while others conveyed personal stories, promoted small businesses, or offered messages for the wider community (see Fig. 7 for examples). The distribution of video purpose categories across the dataset is shown in Table 5, and category definitions and example inputs / outputs are described in Table 6.

6.3.3 Insights from observations and interviews with users and shopkeepers. Enthusiasm, curiosity, and play: Across both photocopy and juice shop deployments we observed clear levels of enthusiasm. People gathered in groups, encouraging one another to experiment with different system modes. Many expressed excitement at seeing themselves, their friends, or local figures “brought to life” in video form. We also observed cases where errors by the system brought amusement. For instance, one user uploaded a photo of an ambulance in a rain drenched street and added a voice prompt, “ambulance flying”. However, the speech recognition service misrecognised the phrase as “ambulance banana”. The result was a surreal video of an ambulance “floating” through a flooded street alongside a man juggling bananas.

The system usage also included many moments of experimentation: for example, one user staged a photo where a person pushed another towards the camera, then reflected aloud that the mixed-up generated faces and clothing must have been due to “[...] two people being in the frame.” Such interpretations became shared learning opportunities, as groups discussed system behaviour.

Demographics and behaviours of users: In manually noting use of the devices at a distance, we found that a majority of users appeared to be younger men (approximately 19–35 years old). We were not present during the entire time the appliances were online, but the automatically-generated descriptions of the uploaded “selfie” type images were also mainly of younger men. We did observe some groups of young women using the device and also more that approached the box out of curiosity but who ultimately walked away without interacting.

Turning to consider the attrition rate logged automatically and reported above, Dharavi is a highly “busy” place and people appear generally to be in hurry to complete tasks. We saw this reflected in their interactions with the shopkeepers when they were buying the services / refreshments, and noted repeated cases where a user would begin to consider the appliance but walked away, especially where there was an interaction step that appeared lengthy (e.g., capturing an image, or the delay as the video was generated).

While we did not provide an automated way for users to take-away videos (see Section 5.1), we observed people using their own

Table 6: Video purpose category definitions. Each generated video was classified into one of these 15 categories. Distributions across the dataset and the prototypes' two input modes are shown in Table 5.

| Category | Description | Example spoken prompt / video |
|-------------------------------|---|--|
| Unclear / exploratory use | Videos where neither the images nor prompts provided a clear purpose, often reflecting experimentation with the system. | No spoken prompt given; video output ambiguous. |
| Explanation of self / objects | Spoken prompts narrated or explained the uploaded image itself. | A student's photo and the prompt <i>"I really enjoy school,"</i> made a cartoon-style video of studying. |
| Creative re-contextualisation | Uploaded images were reimagined through prompts that requested movement, transformations or new elements. | A portrait with the prompt <i>"Make me wrestle a fighter,"</i> ; the resulting video showed the person sparring with a bodybuilder in cartoon style. |
| Positive self-presentation | Videos highlighted personal qualities, often aligning with social media aesthetics. | Prompts like <i>"I am cool,"</i> and videos of participants posed athletically or TikTok / Instagram style. |
| Advertising / promotion | Content used to promote businesses, shops or services. | Shop images paired with descriptive prompts about products for sale. |
| Imagining a better self | Videos where participants placed themselves in aspirational or future roles. | Voice prompts included <i>"Show me living a fulfilled life"</i> |
| Reassurance | Comforting or supportive messages. | Messages such as <i>"All will be OK."</i> |
| Learning | Requests for videos to explain or depict informational content. | Requests such as <i>"Show me the Bihar elections."</i> |
| Romantic themes | Couples shown in sentimental or affectionate contexts. | Prompts and videos depicting romantic poses and interactions. |
| Celebration / greeting | Videos expressing congratulations, joy, or greetings. | Prompts such as <i>"Congratulations on your award."</i> |
| Community assistance | Themes of public information or warnings, often contextualised by local events. | During monsoon flooding, participants generated flood-warning videos for neighbours. |
| Visualising places | Images of locations paired with prompts to transform or reframe them. | For example, <i>"Show me my home village,"</i> or <i>"Show me this place in snow time."</i> |
| Seeking advice / counselling | Videos generated from prompts asking for life guidance or psychological support. | Questions such as <i>"When will I get married?"</i> |
| Family life | Videos of family groups / connections | Wedding shot; <i>"We parents gave life to these children"</i> |
| Futuristic | Stories about far off future/places | Prompts like <i>"Show me a Martian future"</i> |

phones to take pictures / videos of the content. From the Wizard-of-Oz workshops, we might infer that users would subsequently share their videos via social media.

Social and community dynamics: The community showcase that allowed users to browse previously created videos prompted social interactions. In one case, a woman recognised her cousin in a video thumbnail, replaying it repeatedly and drawing in her peers to watch additional videos. Such interactions enabled awareness of the system, discussion and learning, with the appliance becoming a place of gathering.

However, the same browsing feature also generated risks. In one example, a user created a video of a neighbouring shopkeeper with added audio claiming he was located in a local red-light alley. This defamatory video was uploaded to the showcase area, sparking anger and fears about reputational harm. This video and any other similar cases were promptly reported by the shopkeepers and removed from the gallery. Shopkeepers requested a "report" button to give them direct agency in moderating harmful or offensive videos.

The system was also used for social commentary: one older participant attempted to use it to "shame" his adult sons for being lazy by posting a video to the community gallery. Yet the generated output softened his words into a polite fatherly message: he was not happy with the way in which his intentions had been nuanced by the generative AI.

Emerging views on value: While entertainment and experimentation was a dominant use case, several participants explored commercial and aspirational uses. Small business owners experimented with creating advertising content – for example, attempting to superimpose stitched garments onto models or animating business logos. While the system did not always meet expectations (e.g., garments remained flat-lay rather than being "worn"), such attempts highlight the entrepreneurial orientation of some of the local users. Participants in Dharavi's clothing sector, in particular, envisioned the system as a low-cost marketing tool. Others expressed willingness to pay for high-quality short videos akin to

Instagram reels, wedding invites or multilingual promotional clips, suggesting economic possibilities if aligned with local media needs.

Negotiating privacy, trust and misuse: Despite moments of joy, some users voiced concerns around privacy and security. Questions arose such as: “*Will my photo be sent to another country?*” and “*What if my face is used for fraud?*” Given rising fears of impersonation scams,⁷ users worried that the system could be used to fabricate evidence within community disputes.

These anxieties were reinforced by users witnessing direct experiences of misuse of the appliances such as the defamatory shop-keeper video discussed above. While some people resumed use after reassurance, others disengaged entirely, highlighting the importance of trust in such deployments.

Interaction challenges and interface frictions: The hybrid physical-digital nature of the appliance interface produced several usability challenges. Users sometimes mistook the on-screen elements as there to be touched when we had designed physical buttons and dials for those features. For example, users of a touch-screen would attempt to manipulate the style visualisation sliders by touch rather than via the accompanying dials. However, this mismatch did not generate significant frustration – likely because users were already accustomed to similar mixed input modalities in public ATMs (see Fig. 4).

Other issues related to audio input and processing. The device-mounted microphone forced participants to bend down, often competing with roadside traffic noise; some users suggested a gooseneck microphone to improve usability and clarity of captured audio. Participants also highlighted accessibility barriers: speech recognition performed poorly for local dialects. The system’s reliance on “clean Hindi” created moments of breakdown where intended words were substituted with semantically distant alternatives (e.g., “*Dharavi*” misrecognised as “*Dhava*”, meaning medicine).

Desired features and future directions: In addition to those already noted, users offered numerous suggestions for improving future usability and accessibility:

Content control: Report / delete functions for harmful content.

Video sharing: QR codes or downloadable links to share videos beyond the kiosk.

Transparency in prompts: Displaying a representation of the interpreted prompt before generation to help participants refine inputs.

Multilingual interfaces: To support code-switching and regional dialects.

Advanced video production: To support longer video formats and advanced features such as background replacement, video-to-video transformations, and stylised edits.

7 Discussion

The in-situ deployment of the Street Scenes system reveals both the opportunities and the tensions of embedding generative AI in the everyday life of informal settlements.

⁷See, for example, [9], or: <https://indianexpress.com/article/technology/artificial-intelligence/ai-scams-surge-in-india-voice-cloning-deepfakes-and-otp-frauds-leave-victims-helpless-10232064/>

7.1 Portraits as a flexible medium of expression

A striking feature of the deployment data was the number of selfies and portraits among the uploaded images: 67.4% of all 266 images were coded as individual portraits, rising to 75.3% in Bring to life sessions. We see though that while this type of still image was predominant, users used this raw material for diverse forms of self-representation. So, the videos for *Positive self-presentation* (11.3% of all videos) often drew directly from selfies. Participants used voice prompts such as “*I am cool*” or “*Make me a hero*” to transform ordinary photos into stylised, aspirational depictions. The *Creative re-contextualisation* videos (16.5%) also relied heavily on selfies. For example, a participant uploaded their own portrait and requested “*Make me wrestle a fighter*,” producing a surreal video of them sparring with a cartoon bodybuilder. Such uses illustrate how selfies could be re-imagined beyond their original setting, becoming starting points for dramatically distinct end-scenarios. The *Imagining a better self* videos (3.6%) also drew almost exclusively from selfie inputs, with prompts like “*Show me living a fulfilled life*.” These requests placed participants’ likenesses into aspirational contexts, suggesting the appliance’s potential as a tool for projecting identity beyond current circumstances. Even the unclear or exploratory uses (23.2%) often involved selfies, with participants experimenting by uploading their own faces without specifying a clear purpose. By contrast, other categories such as advertising (10.8% of videos) and community assistance (2.1%) were less dependent on selfies, typically drawing instead on images of products, shops or environmental contexts. Taken together, our findings suggest that selfies were not merely the most common input type but also a flexible form through which users explored personal identity, humour, aspiration, and play.

7.2 Engagement as social practice

Far from being a solitary interaction, the appliance became a place of collective engagement and spectatorship. Groups clustered around the appliances, commenting on outputs, laughing at errors and guiding each other’s interactions. The gallery feature amplified this dynamic: browsing others’ creations seeded new participation, with nearly half of all gallery viewers going on to make a video themselves. These patterns position public generative AI not just as a content-production tool but as a street-level media infrastructure – a catalyst for curiosity, gossip, recognition and shared learning. Such shared discovery and learning is vital as AI becomes a powerful, and currently relatively little-known force in informal settlements such as Dharavi. Importantly, however, we also observed uneven participation. While young men dominated direct use, young women showed curiosity but often disengaged before interacting, reflecting broader gendered inequities in digital literacy and public technology use [12]. As noted by Bidwell [4], women’s involvement in community technology projects across the Global South is often constrained by social norms governing mobility, safety and public visibility, which in turn can affect the sustainability of such initiatives. Given the curiosity we observed by women, public appliances may offer an entry point to AI skills and confidence for such under-represented groups, but will require further careful design and facilitation (e.g., situating the appliances in kiosk shops that attract predominately women [8]).

7.3 Risks, misuse and governance

The same affordances that encouraged playful experimentation also created vulnerabilities. There were concerning cases of potential reputational harm, and these incidents highlight the urgency of governance mechanisms beyond technical safety filters. Users themselves suggested features such as report / delete functions, shopkeeper moderation and transparency around how prompts were going to be interpreted. Embedding community-led oversight, rather than only automated safeguards, will be critical for sustaining legitimacy and trust in dense social networks like Dharavi. A practical model to enable this, then, would weave together technical, social and institutional layers of accountability. Shopkeepers—already functioning as trusted gatekeepers—could be equipped with simple moderation tools. A rotating neighbourhood moderation group might then review flagged videos, guided by locally articulated norms of respect, humour and consent. Further, partnerships with local NGOs or municipal digital-literacy programmes could offer light-touch training and audits, ensuring transparency without external control.

7.4 Trajectories of value and sustainability

While our deployment focused on appropriation and social dynamics, the findings also surface clues about how such appliances might be sustained in practice. Participants' behaviours and expressed interests point toward possible economic models. Several users indicated willingness to pay for premium outputs, such as longer or more stylised videos for weddings, celebrations or social media reels. Small business owners experimented with using the system for advertising garments, shops or services, suggesting alignment with local enterprise economies.

The public gallery further revealed the potential of the appliance as a shared media platform, attracting not only creators but also browsers. This audience dimension opens possibilities for sponsorship or light-touch advertising integration (including by NGOs or government providing, for example, public health information), which could help offset computational costs. At the infrastructural level, shopkeepers who hosted the devices saw them as a means of attracting footfall, suggesting that appliances could be sustained through revenue-sharing or rental models, much like photocopy or printing services themselves.

7.5 Prompt scaffolding

Around a quarter of sessions involved a theme button-based prompt. The “Fun” and “Community support” options predominated; and, nearly two-thirds of sessions involved use of the dial controls, showing an appetite to experiment with stylistic parameters such as realism, futurism or shot composition. These interactions indicate participants treated the system not as a “black box” but as a creative instrument, with the physical controls offering a form of scaffolding that enabled experimentation.

7.6 Interaction challenges; design opportunities

While users appeared able to negotiate limitations and interaction breakdowns, these challenges did slow learning and generated confusion. In particular, poor performance of speech recognition for local dialects revealed accessibility exclusions, underscoring the

importance of multilingual, dialect-sensitive and code-switching capable speech recognition modules [11].

Breakdowns also occurred when voice prompts were underspecified or abstract (e.g., “*marriage*”, “*better life*”). These problems highlight the fragility of AI-mediated creativity when input is minimal or ambiguous. Yet ambiguity was also apparent in the interactions we observed and analysed: around a quarter of videos were coded as having unclear purposes. Rather than dismiss these as failures, we suggest them as evidence of a different kind of value: underspecification created space for curiosity and improvisation. Designing to support this balance—scaffolding input quality without closing down exploratory openness—is a key opportunity for future systems.

Users also highlighted demands for a greater match between uploaded images and generated content, options for easier off-appliance sharing, and advanced creative options (longer videos, background replacement, affective tones). These desires show how quickly users may move from appreciating the novelty of a public AI system to expecting professional-grade creative tools. Designing for such aspirational trajectories is as important as accounting for infrastructural constraints.

8 Limitations

While our work offers new insights into the design and deployment of public generative AI appliances in low-resource settings, we acknowledge several limitations, assessing their impact on our findings.

Context and generalisability: Our deployments centred on Dharavi, Mumbai, with prior exploratory workshops in Nairobi. We argue that Dharavi can be seen as a critical case location: one of the largest and most globally recognised informal settlements, with dense social, linguistic and economic diversity. Insights here are unlikely to map one-to-one to other sites, yet the issues we observed—around access, play, governance, and appropriation—are exemplary of those in other lower-resource settings. The Nairobi workshops, while not followed by deployment, provided significant motivations to explore how creative uses of GenAI resonate across geographies.

Prototype fidelity and functionality: Our prototypes relied on commercial AI services with constraints such as limited input formats, short video lengths, and imperfect speech recognition. These limitations shaped user experiences – sometimes frustrating participants; other times prompting playful reinterpretations. While specific model accuracies will improve over time, the structural challenge of noisy, multilingual, and low-literacy environments will remain. Our observations suggest that even with system enhancements, public AI systems must design for breakdowns: providing visible prompt interpretations to help users understand system outputs, offering non-textual modes (e.g., button or dial input) and enabling community members to reinterpret “failed” outputs as moments of humour, play or shared critique.

Study duration: Our in-situ deployments lasted two weeks across two public sites. This period provided rich data on immediate engagement and early appropriation, but does not capture long-term use, sustained community governance or evolving practices of appropriation (cf. [32]).

Ethical considerations and moderation: Despite safeguards such as image filters, the public gallery surfaced risks of reputational harm and misuse, which we mitigated reactively (e.g., by removing defamatory content). Our study did not fully explore systematic governance models or the long-term social consequences of such systems, both of which warrant deeper investigation.

Researcher presence: Finally, the visibility of the research team during deployments may have shaped how people engaged with the appliances – either encouraging experimentation or inhibiting more transgressive uses. While we attempted unobtrusive observation, future work might explore less researcher-mediated deployments.

9 Conclusion

This paper has presented Street Scenes, a public appliance for generative AI video creation, designed, prototyped and deployed in Dharavi, Mumbai. Through a sequence of workshops in Mumbai, India and Nairobi, Kenya, Wizard-of-Oz evaluations, and a 13-day in-situ deployment, we examined how residents of informal settlements engage with generative AI when barriers of cost, literacy, and infrastructure are reduced.

Our findings show that generative AI in public contexts is not simply another information kiosk: it introduces qualitatively new dynamics. Unlike earlier interventions, which often focused on information access or learning, Street Scenes foregrounded creativity, experimentation and authorship. Residents used the appliances to play, to animate their identities, to promote small businesses, and to comment on community life.

At the same time, the deployment surfaced ethical tensions: fears of fraud, misuse for reputational harm and accessibility gaps linked to dialect and literacy. We therefore argue that governance must be designed into public AI appliances from the start. Drawing on community feedback, we outline directions for embedded moderation (e.g., shopkeepers and resident committees), transparent prompting and shared accountability. Such governance shifts AI from a “black box” service into a locally-governed infrastructure.

Taken together, our contributions are threefold:

Conceptual: We advance the notion of public AI appliances as a design space distinct from prior public computing. Here, unpredictability, authorship and governance are defining features rather than side-effects.

Empirical: We provide the first in-situ account of generative AI deployment in an informal settlement, showing how residents appropriate such systems for identity work, affective play, entrepreneurial creativity and community commentary.

Design and governance: We distil actionable lessons for lowering barriers to AI use, scaffolding creativity without closing down ambiguity and embedding moderation and sustainability within existing local economies.

More broadly, our work argues that AI futures in the Global South should not be limited to “worthy” domains like health or education. Play, self-expression and creativity are equally vital to how communities negotiate technology. By extending the lineage from Hole-in-the-Wall to StreetWise to Street Scenes, we show how public AI can be reimagined as situated, collective and accountable infrastructure. We invite the CHI community to take seriously the question of who gets to be creative with AI, under what conditions,

and with whose authority – questions that will only grow more urgent as generative systems become entwined with everyday life worldwide.

Acknowledgments

We thank community members and study participants for their contribution to this research. This work was supported by Engineering and Physical Sciences Research Council grant EP/Y010477/1 and Responsible AI UK grant IA012.

References

- [1] Angela Abascal, Sabine Vanhuyse, Taïs Grippa, Ignacio Rodriguez-Carreño, Stefanos Georganos, Jiong Wang, Monika Kuffer, Pablo Martinez-Diez, Mar Santamaria-Varas, and Eleonore Wolff. 2024. AI perceives like a local: predicting citizen deprivation perception using satellite imagery. *npj Urban Sustainability* 4, 1 (2024), 20. doi:10.1038/s42949-024-00156-x
- [2] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavana, Sidhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling “Culture” in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15763–15784. doi:10.18653/v1/2024.emnlp-main.882
- [3] Elizabeth A Ankrah, Kagonya Awori, Stephanie Nyairo, Mercy Muchai, Millicent Ochieng, Mark Kariuki, Gillian R Hayes, and Jacki O’Neill. 2025. Social by Nature: How Socio-structure Shapes the Work of SMBs and Considerations for Reimagining Collaborative Human-AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI ’25)*. Association for Computing Machinery, New York, NY, USA, Article 1202, 19 pages. doi:10.1145/3706598.3715019
- [4] Nicola J Bidwell. 2020. Wireless in the Weather-world and Community Networks Made to Last. In *Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 1 (Manizales, Colombia) (PDC ’20)*. Association for Computing Machinery, New York, NY, USA, 126–136. doi:10.1145/3385010.3385014
- [5] Nicola J. Bidwell, Masbulele Siya, Gary Marsden, William D. Tucker, M. Tshemese, N. Gaven, S. Ntlangano, Simon Robinson, and Kristen ALI Eglinton. 2013. Walking and the social life of solar charging in rural africa. *ACM Trans. Comput.-Hum. Interact.* 20, 4, Article 22 (Sept. 2013), 33 pages. doi:10.1145/2493524
- [6] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. *Thematic Analysis*. Springer Singapore, Singapore, 843–860. doi:10.1007/978-981-10-5251-4_103
- [7] Jacob T. Browne. 2019. Wizard of Oz Prototyping for Machine Learning Experiences. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI EA ’19)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3290607.3312877
- [8] Dipanjan Chakraborty, Sayonee Chatterjee, Rajeshwari Tripathi, Akshay Gupta, and Seth Seth, Aaditeshwar. 2024. Experiences from Running a Participatory Media Platform for Women and Led by Women in Rural North India. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 194 (April 2024), 23 pages. doi:10.1145/3653685
- [9] Aahana Chopra and Ananya Shukla. 2024. Exploring the Misuse of Deepfake Technology in India: Implications for Society. *Global Media Journal – Indian Edition* 16, 2 (Dec. 2024), 30 pages.
- [10] Andy Crabtree. 2025. H is for human and how (not) to evaluate qualitative research in HCI. *Human-Computer Interaction* 0, 0 (2025), 1–24. doi:10.1080/07370024.2025.2475743
- [11] A. Diwan, R. Vaideeswaran, S. Shah, A. Singh, S. Raghavan, S. Khare, V. Unni, S. Vyas, A. Rajpuria, C. Yarra, A. Mittal, P. K. Ghosh, P. Jyothi, K. Bali, V. Seshadri, S. Sitaram, S. Bharadwaj, J. Nanavati, R. Nanavati, and K. Sankaranarayanan. 2021. MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In *Proceedings of Interspeech 2021*. ICASA, 2446–2450. doi:10.21437/Interspeech.2021-1339
- [12] Cledwyn Fernandez, Havishaye Puri, and Shravani Prakash. 2025. *Digital Technology as an Instrument to Bridge the Gender Gap*. Policy Brief. United Nations Development Programme (UNDP) and Indian Council for Research on International Economic Relations (ICRIER). https://www.undp.org/sites/g/files/zskgke326/files/2025-02/undp-icrier_policy_brief_31.pdf
- [13] David M. Frohlich, Dorothy Rachovides, Kiriaki Riga, Rammath Bhat, Maxine Frank, Eran Edirisinghe, Dhammike Wickramanayaka, Matt Jones, and Will Harwood. 2009. StoryBank: mobile digital storytelling in a development context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*

- (Boston, MA, USA) (*CHI '09*). Association for Computing Machinery, New York, NY, USA, 1761–1770. doi:10.1145/1518701.1518972
- [14] Brett A. Halperin, Diana Flores Ruiz, and Daniela K. Rosner. 2025. Underground AI? Critical Approaches to Generative Cinema through Amateur Filmmaking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1141, 18 pages. doi:10.1145/3706598.3713342
- [15] Mina Huh, Ding Li, Kim Pimmel, Hijung Valentina Shin, Amy Pavel, and Mira Dontcheva. 2025. VideoDiff: Human-AI Video Co-Creation with Alternatives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1143, 19 pages. doi:10.1145/3706598.3713417
- [16] Giulio Iacucci, Kari Kuutti, and Mervi Ranta. 2000. On the move with a magic thing: role playing in concept design of mobile services and devices. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (New York City, New York, USA) (*DIS '00*). Association for Computing Machinery, New York, NY, USA, 193–202. doi:10.1145/347642.347715
- [17] Matt Jones, Dani Kalarikalayil Raju, Jen Pearson, Thomas Reitmaier, Simon Robinson, and Arka Majhi. 2024. Beyond ‘Slumming It’—AI and the Real Lives of Global South Communities. *Interactions* 31, 6 (Oct. 2024), 20–27. doi:10.1145/3696019
- [18] Shivani Kapania, Oliver Sity, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. “Because AI is 100% right and safe”: User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. doi:10.1145/3491102.3517533
- [19] Marc Langheinrich, Albrecht Schmidt, Nigel Davies, and Rui José. 2012. Open Display Networks: A Communications Medium for the 21st Century. *Computer* 45, 05 (May 2012), 58–64. doi:10.1109/MC.2012.114
- [20] Xinrui Lin, Heyan Huang, Kaihuang Huang, Xin Shu, and John Vines. 2025. Seeking Inspiration through Human-LLM Interaction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 150, 17 pages. doi:10.1145/3706598.3713259
- [21] Duri Long, Mikhail Jacob, and Brian Magerko. 2019. Designing Co-Creative AI for Public Spaces. In *Proceedings of the 2019 Conference on Creativity and Cognition* (San Diego, CA, USA) (*C&C '19*). Association for Computing Machinery, New York, NY, USA, 271–284. doi:10.1145/3325480.3325504
- [22] Anunay Mathur, Benjamin Schlotfeldt, and Marshini Chetty. 2015. A Mixed-Methods Study of Mobile Users’ Data Usage Practices in South Africa. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. Association for Computing Machinery, New York, NY, USA, 1209–1220. doi:10.1145/2750858.2807522
- [23] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (Nov. 2019), 23 pages. doi:10.1145/3359174
- [24] Indrani Medhi, Archana Prasad, and Kentaro Toyama. 2007. Optimal audio-visual representations for illiterate users of computers. In *Proceedings of the 16th International Conference on World Wide Web (Banff, Alberta, Canada) (WWW '07)*. Association for Computing Machinery, New York, NY, USA, 873–882. doi:10.1145/1242572.1242690
- [25] Nemanja Memarovic, Sarah Clinch, and Florian Alt. 2015. Understanding Display Blindness in Future Display Deployments. In *Proceedings of the 4th International Symposium on Pervasive Displays* (Saarbruecken, Germany) (*PerDis '15*). Association for Computing Machinery, New York, NY, USA, 7–14. doi:10.1145/2757710.2757719
- [26] Nemanja Memarovic, Marc Langheinrich, Keith Cheverst, Nick Taylor, and Florian Alt. 2013. P-LAYERS – A Layered Framework Addressing the Multifaceted Issues Facing Community-Supporting Public Display Deployments. *ACM Trans. Comput.-Hum. Interact.* 20, 3, Article 17 (July 2013), 34 pages. doi:10.1145/2491500.2491505
- [27] Nusrat Jahan Mim, Dipannita Nandi, Sadaf Sumyia Khan, Arundhuti Dey, and Syed Ishtiaque Ahmed. 2024. In-Between Visuals and Visible: The Impacts of Text-to-Image Generative AI Tools on Digital Image-making Practices in the Global South. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '24*). Association for Computing Machinery, New York, NY, USA, Article 474, 18 pages. doi:10.1145/3613904.3641951
- [28] Sugata Mitra. 2005. Self organising systems for mass computer literacy: Findings from the ‘hole in the wall’ experiments. *International Journal of Development Issues* 4, 1 (2025/07/17 2005), 71–81. doi:10.1108/eb045849
- [29] Dumisani Moyo and Allen Munoriyarwa. 2021. ‘Data must fall’: mobile data pricing, regulatory paralysis and citizen action in South Africa. *Information, Communication & Society* 24, 3 (2021), 365–380. doi:10.1080/1369118X.2020.1864003
- [30] Keziah Naggita, Julienne LaChance, and Alice Xiang. 2023. Flickr Africa: Examining Geo-Diversity in Large-Scale, Human-Centric Visual Data. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) (*AIES '23*). Association for Computing Machinery, New York, NY, USA, 520–530. doi:10.1145/3600211.3604659
- [31] David Nemer. 2022. *Technology of the Oppressed: Inequity and the Digital Mundane in Favelas of Brazil*. The MIT Press. doi:10.7551/mitpress/14122.001.0001
- [32] Stuart Alan Nicholson, Clara Crivellaro, Adrian K Clear, Dan Jackson, Rob Comber, Vasilis Vlachokyriakos, and Patrick Olivier. 2021. Sustaining a Networked Community Resource: Findings from a Longitudinal Situated Display Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI '21*). Association for Computing Machinery, New York, NY, USA, Article 561, 13 pages. doi:10.1145/3411764.3445693
- [33] Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O’Neill. 2024. Beyond Metrics: Evaluating LLMs’ Effectiveness in Culturally Nuanced, Low-Resource Real-World Scenarios. arXiv:2406.00343 [cs.CL]
- [34] Rock Yuren Pang, Hope Schroeder, Kynneddy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 456, 20 pages. doi:10.1145/3706598.3713726
- [35] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, Shashank Ahire, Anirudha Joshi, Deepak Sahoo, Nimish Maravi, and Bhakti Bhikne. 2019. StreetWise: Smart Speakers vs Human Help in Public Slum Settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300326
- [36] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, and Anirudha Joshi. 2019. Diversifying Future-Making Through Iterative Design. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 33 (July 2019), 21 pages. doi:10.1145/3341727
- [37] Rida Qadri, Renee Shelby, Cynthia L. Bennett, and Remi Denton. 2023. AI’s Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) (*FAccT '23*). Association for Computing Machinery, New York, NY, USA, 506–517. doi:10.1145/3593013.3594016
- [38] Ahmad Rahmati, Adam Qian, and Lin Zhong. 2007. Understanding Human-Battery Interaction on Mobile Phones. In *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*. Association for Computing Machinery, New York, NY, USA, 265–272. doi:10.1145/1377999.1378005
- [39] Anjali Raj, Adway Mitra, and Manjira Sinha. 2024. Deep Learning for Slum Mapping in Remote Sensing Images: A Meta-analysis and Review. arXiv:2406.08031 [cs.CV]
- [40] Simon Robinson, Jennifer Pearson, Shashank Ahire, Rini Ahirwar, Bhakti Bhikne, Nimish Maravi, and Matt Jones. 2018. Revisiting “Hole in the Wall” Computing: Private Smart Speakers and Public Slum Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3173574.3174072
- [41] Gloria Ronchi and Claudio Benghi. 2014. Interactive light and sound installation using artificial intelligence. *International Journal of Arts and Technology* 7, 4 (2014), 377–379. doi:10.1504/IJART.2014.066456
- [42] Molly Wright Steenson and Jonathan Donner. 2017. Beyond the personal and private: Modes of mobile phone sharing in urban India. In *The reconstruction of space and time*. Routledge, 231–250. doi:10.4324/9781315134499-11
- [43] Hari Subramonyam, Divy Thakkar, Andrew Ku, Juergen Dieber, and Anoop K. Sinha. 2025. Prototyping with Prompts: Emerging Approaches and Challenges in Generative AI Design for Collaborative Software Teams. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 882, 22 pages. doi:10.1145/3706598.3713166
- [44] Emmanuel Tweneboah, Collins W. Munyendo, and Yixin Zou. 2025. “No, I Can’t Be a Security Personnel on Your Phone”: Security and Privacy Threats From Sharing Infrastructure in Rural Ghana. In *Proceedings of the USENIX Security Symposium*. USENIX, 5131–5148. https://www.usenix.org/system/files/usenixsecurity25-tweneboah.pdf
- [45] Aditya Vashistha, Edward Cutrell, Gaetano Borriello, and William Thies. 2015. Sangeet Swara: A Community-Moderated Voice Forum in Rural India. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (*CHI '15*). Association for Computing Machinery, New York, NY, USA, 417–426. doi:10.1145/2702123.2702191
- [46] Susan P. Wyche and Laura L. Murphy. 2012. “Dead China-make” Phones off the Grid: Investigating Mobile Phone Practices in Rural Africa. In *Proceedings of the Designing Interactive Systems Conference (DIS)*. Association for Computing Machinery, New York, NY, USA, 186–195. doi:10.1145/2317956.2317985
- [47] Susan P. Wyche and Laura L. Murphy. 2013. Powering the Cellphone Revolution: Findings from Mobile Phone Charging Trials in Off-Grid Kenya. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Association

for Computing Machinery, New York, NY, USA, 1959–1968. doi:10.1145/2470654.2466260

- [48] Susan P. Wyche, Sarita Y. Schoenebeck, and Andrea Forte. 2013. "Facebook is a Luxury": An Exploratory Study of Social Media Use in Rural Kenya. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*. Association for Computing Machinery, New York, NY, USA, 33–44. doi:10.1145/2441776.2441783
- [49] Yu Zheng, Hongyuan Su, Jingtao Ding, Depeng Jin, and Yong Li. 2023. Road Planning for Slums via Deep Reinforcement Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Long Beach, CA, USA) (KDD '23)*. Association for Computing Machinery, New York, NY, USA, 5695–5706. doi:10.1145/3580305.3599901

A Wizard of Oz prototype system prompts and configurations

A.1 Prototype 1: initial design

The following prompts refer to the system outlined in [Section 4.1](#).

A.1.1 Configuration A – Voice prompt only.

The following text is a description of a Sora video that someone would like: {description}

Analyse the description and write a very detailed, creative, engaging synopsis for a {video_length}-second video (no sound or commentary). Each scene should be highly detailed.

Return the script in plain text and only describe the visual elements of the video.

Variables and media: The value for {description} is obtained using Google Cloud Speech-to-Text. For the Prototype 1 study, {video_length} was set to 15 (this was due to the narration often being longer than the Sora limit of 20 seconds).

A.1.2 Configuration B – Voice prompt + image upload.

The following text is a description of a Sora video that someone would like: {description}

Next, analyse the description and the following images and write a very detailed, creative, engaging synopsis for a {video_length}-second video (no sound or commentary). Each scene should be highly detailed.

Return the script in plain text and only describe the visual elements of the video.

Variables and media: As in configuration A, the value for {description} is obtained using Google Cloud Speech-to-Text, and the video length was preconfigured. User-provided images are uploaded to the GenAI tool along with the above prompt.

A.1.3 Configuration C – Voice prompt + image upload + buttons / sliders.

{style_text}

The following text is a description of a Sora video that someone would like: {description}

Next, analyse the description and the following images and write a very detailed, creative, engaging synopsis for a {video_length}-second video (no sound or commentary). Each scene should be highly detailed.

Return the script in plain text and only describe the visual elements of the video.

Variables and media: As in configuration B. The value for {description} is created by the following prompt:

The video should reflect the following stylistic qualities: {description_block}.

The thematic content of the video should be guided by: {categories}.

At the top of the output, write a concise, technical paragraph describing the video's visual style, composition, and overall mood and theme

Focus on objective, precise descriptions suitable for an art or film production brief.

Avoid metaphorical or poetic language.

Do not include numeric values or raw style data.

{categories} is a list of the selected category buttons and their meaning. {description_block} describes the data from the sliders using the following algorithm:

```
if score <= 10:
    return f"fully {a} style"
elif score <= 30:
    return f"mostly {a} style with some {b} influence"
elif score <= 49:
    return f"leaning towards {a}"
elif score == 50:
    return f"an even blend of {a} and {b}"
elif score <= 69:
    return f"leaning towards {b}"
elif score <= 89:
    return f"mostly {b} with hints of {a}"
else:
    return f"fully {b} style"
```

{a} refers to the style on the left-hand side of the slider; {b} to the style on the right of the slider. Sliders have a minimum value of 0 and a maximum of 100. For example, the result from using the slider *cartoon* ↔ *photorealistic* with a numerical value of 29:

mostly cartoon style with some photorealistic influence

A.1.4 Configuration D – Image upload + buttons / sliders.

{style_text}

Next, analyse the following images and write a very detailed, creative, engaging synopsis for a {video_length}-second video (no sound or commentary). Each scene should be highly detailed.

Return the script in plain text and only describe the visual elements of the video.

Variables and media: As in configuration C, but without the Speech-to-Text {description} value.

A.1.5 Configuration E – Voice prompt + image upload + buttons / sliders + iteration. This configuration uses the same as configuration C, except that after seeing the generated video, the user was able to refine their voice prompt to improve the generated output.

A.2 Prototype 2: refining the design

The following prompts refer to the system outlined in [Section 4.4](#).

A.2.1 Inspire me – up to 3 images uploaded + voice prompt + buttons / sliders.

```
Write a story based on and including as much detail
as possible from the images sent along with this
text. The story will be used as a prompt to generate
a Sora video {video_length}-seconds in length (no
sound or commentary)
```

```
The video should be styled using this: {style_text}
```

```
Improve the story with the spoken input provided by
the user which is: {description}"
```

Variables and media: The value for {description} is obtained using Google Cloud Speech-to-Text. User-provided images are uploaded to the GenAI tool along with the above prompt. {style_text} is generated by the following prompt:

```
The video should reflect the following stylistic
qualities: {description_block}.
```

```
The theme of the video should be heavily guided by:
{categories}.
```

```
At the top of the output, write a Sora preset
including:
Theme: xxxx
Color: xxxx
Film Stock: xxxx
Lighting: xxxx
Camera: xxxx
```

{categories} and {description_block} are defined as described above.

A.2.2 Bring to life – single image upload + voice prompt + buttons / sliders.

```
Bring the image included here to life by writing a
story that explains what is going on and how the
people in the image might move around and use the
objects. Imagine what is not shown in the image -
whats around the image etc. Use as much detail as
possible from the image to make the story. The story
will be used as a prompt to generate a Sora video
{video_length}-seconds in length (no sound or
commentary)
```

```
Use the first video scene in tandem with this
description: {description}"
```

```
The video should be styled using this description:
{style_text}
```

```
In your response include 'Use the first video scene
in tandem with this description ...' as the
beginning of the text
```

Variables and media: {description} and {style_text} as described above. The user provided image is then uploaded to the GenAI tool using its user interface.

B Deployment system

The following prompts are in regards to the system outlined in [Section 5.2](#)

B.1 Bring to life: single image upload + voice prompt + buttons / dials

```
This image may show a person showing a smartphone
screen, if so, only describe the image shown on the
screen of the smartphone.
```

```
Bring this image to life. Use the following
description as inspiration: {videoDescription}
```

```
Finally, add a narrator's voice to the video, in
Hindi (Do not add this visually)"
```

```
Visual Style Preferences:
{style_description}
```

```
Themes to reflect in the video:
{theme_descriptions}
```

Variables and media: The user provided image is Base64 encoded and uploaded along with this text. {videoDescription} is obtained using Google Cloud Speech-to-Text. {style_description} is created using the following algorithm:

```
enriched_styles = {
  "photorealistic": "Photorealistic (high-detail,
  true-to-life textures, natural lighting)",
  "cartoon": "Cartoon (bold outlines, simplified
  shapes, exaggerated expressions)",
  "old fashioned": "Old fashioned (vintage
  aesthetic, sepia tones, analog textures)",
  "futuristic": "Futuristic (sleek sci-fi
  environments, glowing neon, high-tech
  materials)"
}

angle = angle %

def style(name, weight):
  desc = enriched_styles.get(name, name)
  return f"{int(weight * 100)}%"

if 170 <= angle <= 190:
  a_weight = b_weight = 0.5
elif 0 <= angle < 170:
  b_weight = round(1 - (angle / 170), 2)
  a_weight = round(1 - b_weight, 2)
else: # 190 < angle <= 360
  a_weight = round((angle - 190) / 170, 2)
  b_weight = round(1 - a_weight, 2)

return f"{style(a, a_weight)}, {style(b, b_weight)}"
```

{a} refers to the style when the dial is turned to the left; {b} to the style when the dial is turned to the right. Dials have a minimum value of 0 and a maximum of 360. For example, the result from using the dial *cartoon* ↔ *photorealistic* with a numerical value of 300:

```
65%
```

{theme_descriptions} is created using the following algorithm:

```
themes = [
  ("Idea", "Present a new concept, invention, or
  spark of imagination. Emphasise innovation and
  creativity."),
  ("Community Support", "Promote positive social
  action, volunteering, kindness, or improving
  society."),
  ("Advertisement", "Simulate a promotional video.
  Showcase a product, service, or persuasive
  pitch."),
  ("Fun", "Emphasise humour, joy, entertainment,
  or lighthearted moments. Prioritise energy and
  playfulness.")
]

return [
  desc for i, (theme, desc) in
  enumerate(self.themes)
  if self.button_states[i] == 1
]
```

{button_states} refers to the pressed state of each of the theme buttons. Buttons have the states 0 or 1. For example, the result from pressing the *Idea* and *Advertisement* buttons, would give {button_states}=[1,0,1,0]:

Present a new concept, invention, or spark of imagination. Emphasise innovation and creativity. Simulate a promotional video. Showcase a product, service, or persuasive pitch.

Example:

This image may show a person showing a smartphone screen, if so, only describe the image shown on the screen of the smartphone
Bring this image to life. Use the following description as inspiration: We are the king of Dharavi, how is our name smile?
Finally, add a narrator's voice to the video, in Hindi (Do not add this visually)

Visual Style Preferences:
Style Axis 1:
92%

Style Axis 2:
0%

Style Axis 3:
50%

Themes to reflect in the video:
Emphasise humour, joy, entertainment, or lighthearted moments. Prioritise energy and playfulness.

B.2 Inspire me: up to three images uploaded + voice prompt + buttons / dials

You are a creative storyteller. Create a vivid, engaging story based on the visual content of the provided image descriptions. The story can ignore the order of the images, if it means a better narrative.

The story should be influenced by the following description from the user: {videoDescription}

Focus on characters (so they resemble the person described), settings, and a possible narrative that ties them together. Your output will be concise and suitable for an 8-second video generation prompt.

Also, include descriptions of sound effects that would enhance the scene.

Finally, add a narrator's voice to the video, in Hindi (Do not add this visually)

Themes to reflect in the video:
{theme_descriptions}

Variables and media: {videoDescription} is obtained using Google Cloud Speech-to-Text. {theme_descriptions} is obtained as described above.

The user provided images are uploaded as a Base64-encoded image, to obtain detailed text descriptions to be included at the end of the above prompt:

These images may show a person showing a smartphone screen, if so, only describe the image shown on the screen of the smartphone.
Describe the following images in great detail. The descriptions will be read later by a human. Describe enough so that they can visualise the image.
Describe any people in the image in a way in which an artist could recreate them.
index each image e.g., image 0.....

The above prompt returns a prompt suitable for Veo video generation, to which we append:

Visual Style Preferences:
{style_description}

{style_description} is created as described above.

Example:

Story builder prompt:

You are a creative storyteller. Create a vivid, engaging story based on the visual content of the provided image descriptions. The story can ignore the order of the images, if it means a better narrative The story should be influenced by the following description from the user: Write about marriage, my dream is to fulfill someones wish. Focus on characters (so they resemble the person described), settings, and a possible narrative that ties them together. Your output will be concise and suitable for an 8-second video generation prompt. Also, include descriptions of sound effects that would enhance the scene.Finally, add a narrator's voice to the video, in Hindi (Do not add this visually).

Themes to reflect in the video:
Promote positive social action, volunteering, kindness, or improving society.

image 0:
This image captures the front view of a dark teal or petrol blue BMW sedan, prominently featuring its distinctive grille. The car is parked on what appears to be a paved surface, possibly a street or a parking area, with a slight curb visible in the foreground.
.....

Veo prompt:

****Visuals for 8-second video:****

****0-3 seconds:**** A close-up on Aryan's face. His dark, expressive eyes, framed by prominent eyebrows, gaze directly at the viewer with a thoughtful, almost questioning intensity. A faint mustache is visible above his subtly closed lips. His dark, voluminous hair is neatly swept back.

****3-8 seconds:**** The camera slowly pulls back to reveal Aryan's full figure. He stands casually in a black long-sleeved top and dark, distressed jeans, one hand tucked into his pocket. The background blurs into a warm, inviting mosaic of soft, multi-colored lights (pink, purple, yellow hues), indistinct figures, and stacked wooden chairs, suggesting a lively yet intimate cafe or event space. His direct gaze holds, hinting at a quiet introspection amidst the gentle hum of the world around him, as if silently posing the question: "How are you?"

****Sound Effects:****

```
* Soft, distant murmur of cafe chatter.  
* Subtle, occasional clinking of glasses and  
cutlery.  
* Warm, muted ambient music, like soft lounge jazz.  
  
**Narrator's Voice (Hindi - Not visual):**  
(Translation: He paused for a moment, his eyes  
asking some deep questions.)  
  
Visual Style Preferences:  
Style Axis 1:  
0%  
  
Style Axis 2:  
7%  
  
Style Axis 3:  
50%
```