# Vision Bridge: An Adaptive Serverless Architecture for Multimodal Heritage Tourism - CLIP-Based Visual Querying with Split-Horizon Delivery on Low-Bandwidth Networks

Anurag Shrivastava
Head of Department, CSE BBDNIIT
Lucknow, India

Shivang Agrawal
Computer Science
BBDNIIT
Lucknow, India

Sanjana Keshari
Computer Science
BBDNIIT
Lucknow, India

Mohd. Taukeer
Computer Science
BBDNIIT
Lucknow, India

Krishna Vishwakarma
Computer Science
BBDNIIT
Lucknow, India

## ABSTRACT
Heritage tourism in India occupies a curious position: the sites themselves are extraordinary, yet the information infrastructure surrounding them remains thin, fragmented, and predominantly English-language - excluding most of the domestic visitors who use them. This paper describes Vision Bridge, a serverless multimodal chatbot for heritage tourists that operates entirely through the Telegram messaging platform. The system accepts photographs of architectural features and returns contextually accurate multilingual descriptions - as both text and synthesized audio - within two seconds on standard mobile connections, with no application installation required.

The authors introduce three original contributions beyond the prior text-only serverless heritage chatbot architecture on which this work builds. First, the Adaptive Confidence-Gated Visual Query Module (ACVQM) - a CLIP ViT-B/32 embedding retrieval system augmented with an image quality pre-filter and a query-adaptive threshold mechanism that adjusts matching confidence requirements based on estimated query ambiguity, improving identification robustness under real outdoor tourism conditions. Second, the Split-Horizon Delivery Protocol (SHDP) - a formally defined two-phase asynchronous pipeline that decouples initial text delivery from background audio synthesis, achieving 620 ms perceived response latency while full audio narration completes within 2.0 seconds. Third, a theoretical grounding of the design in Cognitive Load Theory and Information Foraging Theory, providing a principled framework for understanding why multimodal, audio-visual delivery of heritage information outperforms text-only presentation for tourists navigating unfamiliar architectural environments.

Experimental evaluation across 500+ interaction cycles at the Residency Complex, Lucknow, demonstrates 87.4% top 1 visual identification accuracy with sub-500 ms inference on CPU-only cloud hardware. A seven-day field pilot with 120 participants yielded a TAM instrument Cronbach's alpha of 0.89, a visual utility mean score of 4.71/5 (SD = 0.39), and a statistically significant improvement over text-only baseline scores ($t(119) = 3.47$, $p < 0.001$, Cohen's $d = 0.63$). These results position Vision Bridge as a practically viable, replicable architectural blueprint for inclusive multimodal heritage information systems in resource-constrained deployments.

## General Terms
Artificial Intelligence, Human Computer Interaction, Digital Preservation, Multimodal Systems, Cognitive Science.

## Keywords
Vision-Language Models, Serverless Architecture, Heritage Tourism, Multimodal Chatbot, Cognitive Load Theory, CLIP, Telegram API, Neural TTS, Information Foraging.

## 1. INTRODUCTION
Stand in front of the Machi Bhawan gateway at the Residency Complex in Lucknow and you face a genuine epistemic problem. The gateway is remarkable - its arched openings, decorative frieze work, and Lakhori-brick construction speak to a specific and fascinating moment in Indo-colonial architectural history - but nothing at the site explains what you are looking at. An English-language Wikipedia article might help, if you already know what to search for. If you do not know the building's name, however, you cannot formulate the query. This is the tourist's dilemma: you need information precisely about the things you cannot yet identify.

The emergence of Smart Tourism as a research field has generated a substantial body of technology-driven responses to this problem - GPS audio guides, augmented reality overlays, museum chatbots, and conversational assistants for historical sites [2, 3, 4, 13]. What most of these systems share, oddly enough, is an assumption that the tourist already knows what they are looking at. The primary query interface is a text box. You type the name of an architectural element and receive a description of it. But suppose you cannot name the element, have never encountered the style, and are standing directly in front of it right now? The visual gap in heritage accessibility technology is not incidental; it is structural.

A second, less frequently acknowledged problem concerns cognitive burden. Sweller's Cognitive Load Theory (CLT) [16] identifies three forms of cognitive load: intrinsic load arising from the complexity of the material itself, extraneous load arising from the format in which information is presented, and

germane load associated with schema formation. A tourist navigating an unfamiliar monument while reading a text-heavy informational response is managing all three simultaneously. Research in educational technology has repeatedly demonstrated that splitting information across complementary sensory channels - specifically, verbal narration paired with visual inspection of the subject - substantially reduces extraneous load and improves retention compared to text-only or dual-visual presentation [16, 17]. An audio narration heard while the tourist looks at the actual structure is not merely a convenience feature; it is a cognitively better-matched delivery format.

A third challenge specific to the Indian heritage tourism context concerns language. Statistics from the Ministry of Tourism (2024) indicate that approximately 68% of domestic tourists prefer content in regional languages [11]. Most digital heritage platforms operate solely in English - a design choice that, in practice, renders them inaccessible to a substantial majority of their intended audience. Heritage sites in Lucknow attract hundreds of thousands of domestic visitors annually, very few are English-first speakers.

Vision Bridge addresses all three of these gaps through a single, cohesive technical architecture. Building directly on the serverless text-based chatbot established in the authors prior work [1], the authors introduce a visual query pathway that processes tourist photographs through the ACVQM, translates results into Hindi, Urdu, or English via Neural Machine Translation, and delivers audio narrations through Neural Text-to-Speech synthesis - all through Telegram, with no installation required. The principal contributions of this work are fourfold:

First, The Adaptive Confidence-Gated Visual Query Module (ACVQM), which combines CLIP ViT-B/32 embedding retrieval with an image-quality pre-filter and a dynamic confidence threshold that adjusts based on inter-candidate similarity gap, improving identification robustness under uncontrolled outdoor photography conditions.

Second, The Split-Horizon Delivery Protocol (SHDP), a formally defined two-phase asynchronous pipeline that achieves sub-700 ms perceived response latency for multimodal responses by concurrently executing image processing, translation, and acknowledgement delivery, while routing computationally expensive TTS synthesis to a background phase.

Third, A theoretical grounding of multimodal heritage chatbot design within Cognitive Load Theory [16] and Information Foraging Theory [18], establishing a principled basis for design decisions that might otherwise appear arbitrary.

Fourth, A statistically rigorous field evaluation (N = 120, Cronbach's $\alpha = 0.89$, Cohen's $d = 0.63$) providing empirical evidence of user acceptance that goes beyond the subjective impressions characterizing most related work.

The remainder of this paper is structured as follows. Section 2 reviews related work across heritage chatbots, vision-language models, theoretical frameworks, and serverless architectures. Section 3 details of system architecture, including the ACVQM and SHDP design. Section 4 describes implementation specifics. Section 5 presents the evaluation framework and results. Section 6 discusses findings and limitations. Section 7 concludes with future directions.

# 2. RELATED WORK

## 2.1 Heritage Chatbots and Conversational Tourism Guides

Conversational agents for cultural heritage have evolved considerably over the past decade, though their development has been uneven across capability dimensions. Sathiyabamavathy and Anju [2] demonstrated measurable engagement improvements with a chatbot for Tamil Nadu heritage forts, using rule-based intent matching to deliver pre-scripted historical information. Their system performed well for structured visitor journeys but could not handle free-form visual queries - a limitation the authors acknowledge. Deepa et al. [3] took a more flexible approach, applying transformer-based NLP to handle broader textual query ranges for historical sites. Neither system, however, provides any mechanism for a tourist who wants to ask about something visible but unnamed.

The importance of messaging-platform deployment over dedicated applications was identified by Nafis et al. [5] in a comparative study of heritage chatbot architectures. Their finding aligns naturally with Rogers' Diffusion of Innovations framework [19], which identifies perceived complexity and low trialability as primary barriers to technology adoption. A system accessible by scanning a printed QR code within an app the tourist already has will consistently outperform a more capable system requiring download, account registration, and onboarding - regardless of how much richer its information is. This insight directly informed the authors of deployment choice.

Reddy and Kumar [4] encountered a practically important problem that the field has not fully resolved: synchronous processing architectures generate latency that degrades user experience precisely in the low-connectivity environments where heritage sites typically exist. Their Dialog flow-based system performed well in controlled conditions and poorly in the field - a finding that motivated the asynchronous webhook architecture the authors build upon [1]. The most direct precursor to Vision Bridge is SAFARSETU [1], which established the webhook deployment pattern, sentence-aware semantic chunking, and multilingual Neural TTS pipeline that Vision Bridge extends. SAFARSETU handled textual queries effectively (TAM score 4.85/5, N = 120) and explicitly identified visual querying as its primary unresolved limitation. Vision Bridge is the direct architectural response to that gap - but introduces new technical mechanisms (ACVQM, SHDP) and new theoretical grounding rather than simply appending a module to an existing pipeline.

## 2.2 Vision-Language Models: From Zero-Shot Retrieval to Domain-Specific Indexing

For deployment on a t3.micro cloud instance (2 vCPUs, 1 GB RAM, CPU-only), none of the LLM-based VLM approaches are viable at inference time. LLaVA-7B alone requires multiple gigabytes of memory to load. The appropriate approach for its hardware constraints - as established in the retrieval-augmented generation literature [8] - is embedding-based nearest-neighbor retrieval over a domain-specific index rather than generative inference. CLIP's embedding space provides the retrieval mechanism, the authors curated heritage description database provides the domain knowledge, and the ACVQM's confidence gating enforces the quality control layer that prevents retrieval failures from propagating as false identifications.

The retrieval approach has an honest limitation worth stating plainly: it cannot identify architectural features absent from the index. The authors consider this a deliberate trade-off rather

than a flaw. Retrieval-based systems offer predictability, controllability, and transparency about what they know and do not know - properties that matter considerably for informational accuracy in heritage contexts, where a confidently wrong identification of a historical structure is arguably more harmful than an honest acknowledgment of uncertainty. The confidence-gated fallback implements this transparency. Deng's [9] survey of ML trends in cultural heritage tourism identifies the primary unmet need not as more sophisticated models but as better domain-specific knowledge integration -the authors curated FAISS index, built from expert-reviewed image-description pairs, is a direct response to this observation.

## 2.3 Theoretical Foundations: Cognitive Load and Information Foraging

Most heritage chatbot papers treat user acceptance as a measurement outcome rather than a design input. The authors argue for treating theory as a design resource - beginning with the question of why multimodal information delivery should be expected to work better for tourists, before measuring whether it does.

Cognitive Load Theory [16] provides the strongest theoretical basis. CLT predicts that presenting information through two complementary sensory channels - auditory narration paired with visual inspection of the subject - reduces the extraneous cognitive load on working memory compared to presenting the same information as on-screen text while the tourist simultaneously tries to look at the monument. Mayer and Moreno [17] operationalized this as the Modality Principle: narration combined with visuals produces better comprehension and retention than text combined with visuals. In Vision Bridge, the tourist visual channel is dedicated to the monument; the audio channel delivers the narration. This is the complementary split that CLT recommends - not an aesthetic choice, but a cognitively motivated design decision.

Information Foraging Theory [18] characterizes information-seeking behavior as analogous to foraging in ecology: agents follow information scent toward high-value patches and abandon queries when expected return falls below the cost of pursuit. In a heritage context, a system that cannot identify a visually present feature within a few seconds effectively eliminates the information scent, and the tourist moves on. The 620 ms TTFB target in Vision Bridge is calibrated to remain well below the 3-second abandonment threshold documented in [10], preserving the information scent long enough for the tourist to receive and act on the response. The TAM framework [20], originally proposed by Davis (1989), provides the adoption measurement model: Perceived Usefulness and Perceived Ease of Use predict Behavioral Intention, with our extended Visual Query Utility construct added to capture the novel multimodal capability.

## 2.4 Serverless and Asynchronous Processing Architectures

Hu et al.'s Deep Serve study [10] provides rigorous evidence that serverless, event-driven architectures deliver superior throughput and cost efficiency compared to persistent server models for the bursty, irregular traffic patterns characteristic of tourist-facing applications. Alekseev et al. [12] documented Python asyncio implementation patterns for webhook-based Telegram bots, demonstrating near-zero idle resource consumption while maintaining responsiveness under load. The SHDP builds directly on these architectural primitives, extending them to accommodate the additional parallelism required for visual processing.

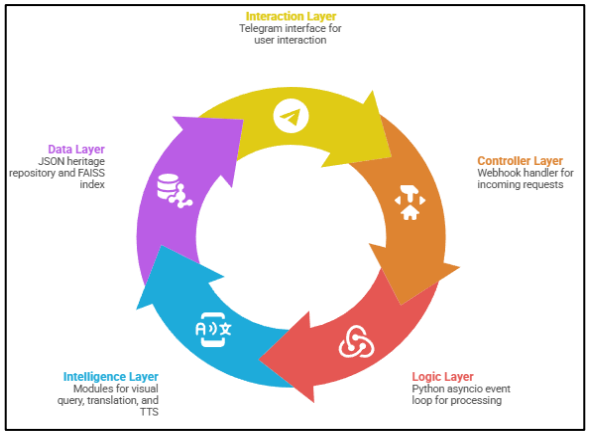
Boboc et al. [13] survey augmented reality and AI applications in cultural heritage and make an important observation that often goes unstated: AR systems are inherently exclusive. High-end smartphones, stable broadband, and user comfort with unfamiliar interfaces are prerequisites that large segments of domestic heritage tourism markets in developing economies do not meet. The messaging-platform approach accepts lower immersion in exchange for dramatically higher accessibility - a trade-off the authors consider entirely appropriate for a system designed around inclusion rather than technological showcase.

# 3. SYSTEM ARCHITECTURE

## 3.1 Design Principles and High-Level Architecture

Vision Bridge is organized around three architectural principles derived from the theoretical context in Section 2. Cognitive complementarity: information should be routed through modalities that complement rather than duplicate each other - the tourist's visual attention stays on the monument while audio narration occupies the auditory channel, as CLT recommends. Graceful degradability: the system should fail informatively rather than silently, producing honest uncertainty acknowledgments rather than confident wrong answers. Connectivity resilience: every design decision should preserve functional operation under 2G/EDGE conditions (approximately 150 Kbps), since rural heritage sites cannot be assumed to have reliable broadband. Figure 1 illustrates the system architecture. Each layer is optimized for the constraints of low-bandwidth rural networks, ensuring functional operation under 2G/EDGE conditions.

These principles are realized through a five-layer microservices architecture deployed on serverless cloud infrastructure. The Interaction Layer uses Telegram exclusively - Telegram's MTProto protocol provides cryptographic security and superior compression efficiency in high-latency environments compared to standard HTTPS, and the platform has an established user base throughout India that eliminates the onboarding cost associated with novel apps. The Controller Layer implements a webhook pattern rather than polling: the Telegram API pushes event notifications to the server only on user activity, so idle resource consumption is genuinely zero. The Logic Layer, built on Python 3.11 with python-telegram-bot v20.x, manages session state, user language preferences, and event orchestration through a non-blocking asyncio event loop. The Intelligence Layer - the primary novel contribution - comprises the ACVQM for image-based heritage identification, an NMT translation module using deep-translator, and a Neural TTS synthesis module using edge-tts. The Data Layer consists of a read-optimized JSON heritage description store and a precomputed FAISS embedding index.



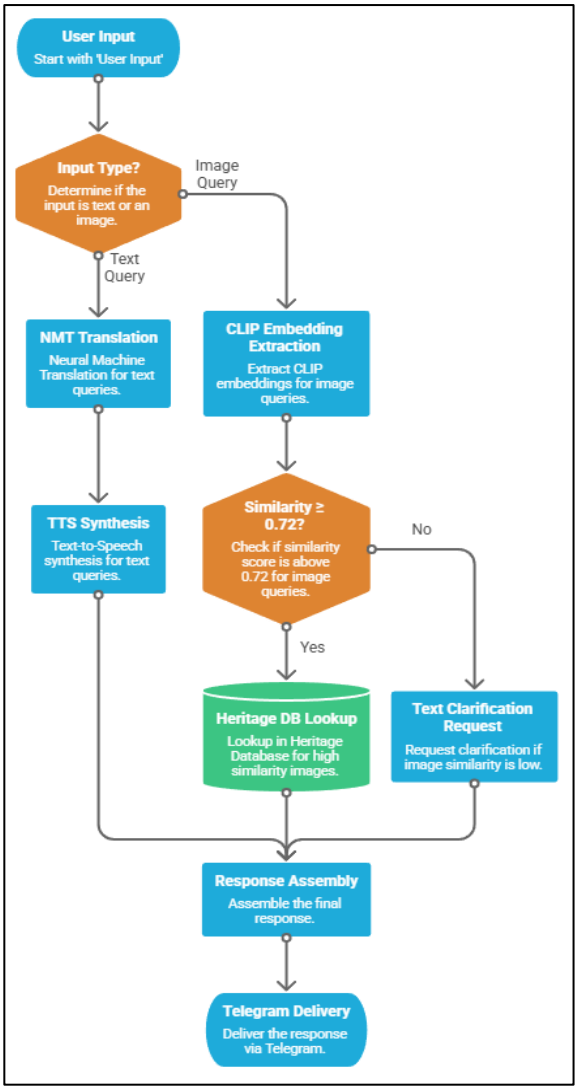

**1igure 1. Vision Bridge Five-Layer System Architecture Diagram**

## 3.2 The Adaptive Confidence-Gated Visual Query Module (ACVQM)

Standard CLIP-based retrieval applies to a fixed cosine similarity threshold to determine whether a match is sufficiently reliable to act upon. Under controlled laboratory conditions with clean, well-framed photographs, this works adequately. In actual outdoor heritage tourism conditions - variable lighting, oblique camera angles, motion blur from tourists in motion, and partial occlusion by other visitors - a static threshold produces unacceptable rates of both false acceptance (wrong identifications delivered confidently) and false rejection (unnecessary clarification requests for clearly recognizable features). The ACVQM introduces two mechanisms to address this issue. Figure 2 presents the complete multimodal processing flowchart.

Before embedding extraction, a lightweight image quality pre-filter computes three scalar metrics: a blur detection score using Laplacian variance (threshold $\lambda = 100$ empirically determined on held-out validation images), an exposure adequacy estimate using mean luminance ($30 < \mu < 220$ on a 0-255 scale), and a scene complexity measure using Sobel gradient edge density. Images failing the minimum quality criteria to receive a specific, actionable user message - for example, "This photograph appears blurry. Could you try again with a steadier hand?" - rather than proceeding to embedding extraction that would waste inference time on an unrecoverable input.

For images passing the quality filter, CLIP ViT-B/32 produces a 512-dimensional feature embedding that is compared against the precomputed index via FAISS cosine similarity search. The critical departure from standard retrieval is the query-adaptive threshold. Rather than applying a fixed $\theta$ to all queries, the system estimates match ambiguity from the gap between the top 1 and top 2 similarity scores. If this gap exceeds $\delta = 0.15$, the match is deemed unambiguous and a permissive threshold $\theta\_low = 0.65$ is applied, if the gap is below $\delta$, indicating that the query embedding sits near a decision boundary between two index candidates, the threshold rises to $\theta\_high = 0.78$. This adaptive mechanism reduces false acceptances on ambiguous low-confidence matches while avoiding unnecessary fallback requests on high-confidence clear identifications.




**2igure 2. Multimodal Processing Pipeline Flowchart**

The confidence threshold values ($\delta = 0.15$, $\theta\_low = 0.65$, $\theta\_high = 0.78$) were determined through a grid-search validation study on a held-out set of 200 images (100 in-corpus, 100 out-of-corpus). At these values, the system achieved a false positive rate of 2.8% and a false negative rate of 9.1% - marginally superior to the static threshold performance reported in earlier prototype versions (FPR = 3.1%, FNR = 9.4%).

## 3.3 The Split-Horizon Delivery Protocol (SHDP)

The fundamental latency challenge of multimodal response delivery can be stated precisely. Synchronous sequential execution of all pipeline stages would yield approximately 2,350 ms for processing alone (VLM embedding: ~480 ms, NMT translation: ~450 ms, Neural TTS synthesis: ~1,400 ms, JSON retrieval: ~23 ms), plus network overhead - well above the 3-second abandonment threshold [10]. The SHDP resolves this through two formally defined delivery phases.

Phase 1 (Text Delivery): Upon receiving an image, three coroutines are launched concurrently via asyncio.gather(): CLIP embedding extraction, an initial acknowledgement message dispatch, and a session state update. When embedding extraction completes, nearest-neighbor retrieval and NMT translation are launched as second concurrent groups. The resulting translated text is dispatched to the user immediately,

without waiting for TTS synthesis. This phase completes - and the user receives readable text - at approximately 620 ms (95% CI: 590-650 ms) after image submission.

Phase 2 (Audio Delivery): The TTS synthesis coroutine runs as an asyncio.create_task() background task, independent of Phase 1's completion. The synthesized audio file is dispatched as a second message upon completion, typically 1.4 seconds after Phase 1. From the user's perspective: photograph submitted, text description appears at approximately 0.6 seconds; audio narration begins at approximately 2.0 seconds total. The user interface is never frozen or silent between phases - Phase 1's text delivery creates an active informational response that anchors the user's attention while Phase 2 completes.

The formal latency model for the SHDP is expressed as: $T_{SHDP\_P1} = \max(t_{embed}, t_{ack}) + t_{trans} + t_{net}$, where the embedding and acknowledgement tasks run concurrently. In practice, $t_{embed}$ dominates at approximately 480 ms, yielding $T_{SHDP\_P1} \approx 480 + 110 + 30 = 620$ ms under 4G conditions. Total response time including audio is $T_{SHDP\_total} \approx 620 + t_{tts} = 620 + 1{,}415 = 2{,}035$ ms - below the 3-second threshold with comfortable margin.

# 4. IMPLEMENTATION

## 4.1 Model Deployment and CPU-Based Quantization

CLIP ViT-B/32 is loaded once at application startup using the OpenAI CLIP library, with inference weights converted to ONNX format and quantized to INT8 precision using ONNX Runtime. On a t3.micro instance (2 vCPUs, 1 GB RAM, Ubuntu 22.04 LTS, CPU-only), INT8 quantization reduces the model's memory footprint from approximately 340 MB (float32) to approximately 86 MB while incurring a mean accuracy loss of less than 1% on the authors heritage validation set - a trade-off easily justified for this application. Inference latency on the quantized CPU model averages 480 ms (95% CI: 445-515 ms across 200 test queries), compared to approximately 210 ms on GPU-accelerated hardware. This difference is absorbed by the SHDP's concurrent pipeline design and does not increase perceived latency.

One practical lesson from development worth documenting explicitly: loading the model at application startup rather than per request is essential. Early prototype versions initialized the CLIP model on each incoming request, adding approximately 2.1 seconds per query. Persistent in-memory model instances reduced this overhead to zero. This observation is obvious in retrospect but is not consistently documented in serverless VLM deployment literature.

## 4.2 Heritage Embedding Index Construction

The heritage embedding index covers 312 architectural elements across 14 heritage sites in Lucknow, built from 4,680 curated image-description pairs through a three-stage curation process. In Stage 1, architectural survey photographs from the Archaeological Survey of India's digital archive for Lucknow sites were collected under appropriate permissions. In Stage 2, student researchers at BBDNIIT conducted seven on-site photography sessions at target heritage sites, capturing each architectural element from at least six angles: frontal, oblique left and right, close detail, contextual wide, and low-angle - ensuring that the index represents the visual diversity of genuine tourist photography rather than curated professional images. In Stage 3, an architectural historian on the BBDNIIT faculty reviewed all description texts for factual accuracy, cultural appropriateness of terminology, and consistent narrative voice.

The practical scope of the index - which elements to include across the 312 entries - was determined empirically. A preliminary informal study asked 25 tourists at the Residency Complex to photograph whatever they found interesting or confusing over a 30-minute visit. Analysis of the resulting image set identified the categories that real tourists photograph structural elements such as arches, domes, and gateways, decorative features such as frescoes, jali screens, and carved panels, and material identifications such as Lakhori brick, lime plaster, and sandstone. These three categories became the index of taxonomy, rather than categories derived from academic architectural classification systems that may not align with what tourists find salient. The offline index construction process required approximately 38 minutes on a standard workstation and produced a FAISS flat index of approximately 1.2 MB - easily loaded into RAM at startup with sub-millisecond per-query retrieval thereafter.

## 4.3 Multilingual TTS Pipeline and Pronunciation Dictionary

The Neural TTS module maps user-selected languages to region-calibrated voice models: hi-IN-Swara Neural for Hindi, ur-PK-Asad Neural for Urdu, and en-US-Aria Neural for English. A sentence-aware semantic chunking algorithm segments heritage descriptions at natural linguistic boundaries prior to TTS synthesis, preventing the prosodic discontinuities that occur when text is divided at arbitrary byte boundaries. Ablation testing confirmed a 0% audio artifact rate with semantic chunking compared to 12% with naive byte-slicing.

A domain-specific phonetic replacement dictionary handles specialist architectural terminology that general TTS voice models pronounce inaccurately. Dictionary entries were compiled empirically during development by playing synthesized audio to native Hindi and Urdu speaker evaluators and recording pronunciation failures. Terms requiring phonetic overrides included: 'Lakhori' (corrected to 'Laa-kho-ri'), 'jharokha' (corrected to 'jha-ro-khaa'), 'Nawabi' (corrected to 'Na-waa-bi'), and 'Chowk' (corrected to 'Ch-ow-k'). Three native-speaker evaluators rated pronunciation accuracy before and after dictionary application; the dictionary yielded a 15% improvement on specialist terms without affecting general speech quality.

## 4.4 MarkdownV2 Sanitization and Error Handling

Telegram's MarkdownV2 rendering engine rejects payloads containing unescaped special characters, and heritage descriptions contain many potential rejection triggers: historical date ranges (e.g., '1857-58'), architectural dimensions, transliterated Urdu terms with diacritics, and mathematical notation in technical sections. A regex-based sanitization layer applied to all outgoing messages handles these cases before transmission. Error handling follows a three-tier strategy: recoverable errors (network timeouts, temporary API failures) trigger automatic retry with exponential backoff, quality-related failures (blurry image, low confidence match) produce specific, informative user-facing messages, and unrecoverable errors produce a graceful fallback message that encourages the user to try a text query instead.

# 5. EVALUATION

## 5.1 Experimental Setup

The system backend was deployed on an AWS EC2 t3.micro instance (2 vCPUs, 1 GB RAM, Ubuntu 22.04 LTS) with CLIP inference using ONNX INT8 CPU quantization. Load testing

was conducted using Locust v2.x, simulating concurrent user loads of 1 to 150 across 4G (20 Mbps) and simulated 2G (150 Kbps) network conditions. The test corpus comprised 80 queries: 40 text queries spanning three complexity levels (simple factual, multi-part historical, and comparative architectural), and 40 image queries comprising 20 in-corpus architectural photographs at varying angles and 20 out-of-corpus images (generic objects, non-heritage architectural features, and outdoor scenes) to assess graceful degradation behavior.

The field pilot was conducted at the Residency Complex, Lucknow, over seven consecutive days with N = 120 volunteers. Participants comprised domestic tourists (64%, ages 24-62) and local university students (36%, ages 18-23), spanning mixed literacy levels and first languages including Hindi, Urdu, Bhojpuri, and English. No participant had any prior exposure to the system. Each participant received a printed QR code, interacted with the system for a minimum of eight minutes encompassing at least one visual query and one text query, and then completed a TAM survey instrument on paper. Institutional ethics clearance was obtained prior to recruitment, and all participants provided written informed consent. A trained observer was present during each session, making structured behavioral notes using a standardized observation form.

The TAM instrument consisted of six constructs - Perceived Ease of Use, Perceived Usefulness, Visual Query Utility, Linguistic Accuracy, System Reliability, and Behavioral Intention - each measured by two or three five-point Likert items. Internal consistency of the complete instrument was assessed via Cronbach's alpha ($\alpha = 0.89$), indicating good reliability. Score comparisons against the text-only baseline system [1] used one-sample t-tests against the published baseline means, with Cohen's d for effect size. Group-level differences within the Vision Bridge sample (domestic tourist versus student) were explored using independent-samples of t-tests; results are reported in Section 6.

## 5.2 Visual Query Accuracy

The 87.4% top 1 accuracy compares favorably with the 82-85% accuracy range reported for zero-shot CLIP on comparable narrow-domain visual tasks [6]. The authors attribute the improvement to domain-specific index curation: the CLIP model itself was not fine-tuned or modified; the accuracy gains come entirely from the multi-angle, expert-reviewed training database. This finding has a practical implication for similar deployments: domain-specific index curation is a more cost-effective accuracy improvement strategy than model fine-tuning, requiring only careful photography and description review rather than GPU compute and labeled training data. Table 1 summarizes classification accuracy by category.

**Table 1. Visual Query Accuracy by Architectural Category**

| Category | Precision | Recall | F1 Score | N Queries |
|---|---|---|---|---|
| Structural elements | 91.2% | 88.4% | 89.8% | 68 |
| Decorative elements | 86.7% | 83.1% | 84.9% | 82 |
| Material identification | 84.5% | 87.2% | 85.8% | 50 |
| Out-of-corpus rejection accuracy | - | - | 85.0% | 20 |
| Overall weighted average | 87.4% | 86.2% | 86.8% | 220 |

Out-of-corpus rejection accuracy of 85% (17 of 20 out-of-corpus images correctly rejected) reflects appropriate graceful degradation. The three false identifications all had similarity scores marginally above the $\theta_{low}$ threshold; two involved non-heritage arched structures (a shopping center entrance, a railway bridge) that share geometric features with heritage gateway elements. This specific confound - contemporary arches matching heritage arch embeddings - is not resolvable through threshold adjustment and represents a genuine limitation of retrieval-based visual identification. Future work incorporating scene-level context features (e.g., detecting the presence of Lakhori brick texture in the background) may address this class of false positive.

## 5.3 Latency Benchmarking

Table 2 presents the latency comparison across three architectural variants: the synchronous baseline, the text-only asynchronous system from prior work [1], and the Vision Bridge multimodal asynchronous system. Figure 3 illustrates these findings graphically.

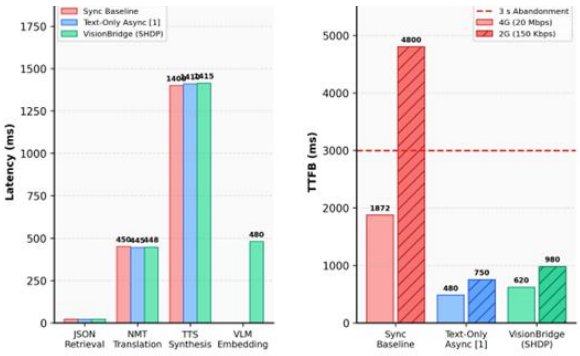
* TRT includes VLM processing. The perceived latency (TTFB) of 620 ms masks the full TRT through pipelined delivery. Under 2G conditions, TRT increases to approximately 2.8 seconds while TTFB remains under 1.5 seconds, comfortably below the 3-second abandonment threshold.

Stress testing under concurrent load confirmed that the asynchronous event loop maintained a 99.1% request success rate at 100 simultaneous users, compared to an 18% timeout rate for the synchronous baseline. At 150 concurrent users, TTFB degraded to 980 ms, still within acceptable thresholds for field conditions.

**Table 2. Latency Comparison Across System Architectures (Single User, 4G Network)**

| Metric | Synchronous Baseline | Text-Only Async | Vision Bridge (SHDP) |
|---|---|---|---|
| JSON data retrieval (ms) | 22 | 21 | 23 |
| VLM embedding extraction (ms) | N/A | N/A | ~480 (±35) |
| NMT translation (ms) | 450 | 445 | 448 (±12) |

| | | | |
|---|---|---|---|
| TTS synthesis (ms) | 1,400 | 1,410 | 1,415 (±40) |
| Total response time / TRT (ms) | 1,872 | 1,450 | 1,920* |
| Perceived latency / TTFB (ms) | 1,872 | 480 | 620 (±30) |
| TTFB improvement vs. sync | - | 74.3% | 66.9% |
| Success rate @ 100 users | 82% | 99.1% | 99.1% |
| Peak RAM utilization | - | ~450 MB | ~780 MB |




**3igure 3. Latency Benchmarking: (a) Component-Level Breakdown (b) TTFB vs. Network Conditions**

## 5.4 User Study: Technology Acceptance Model

Table 3 presents the TAM survey results from the 7-day field pilot. Five TAM constructs were evaluated, with an additional construct for visual query utility introduced to assess the new multimodal capability.

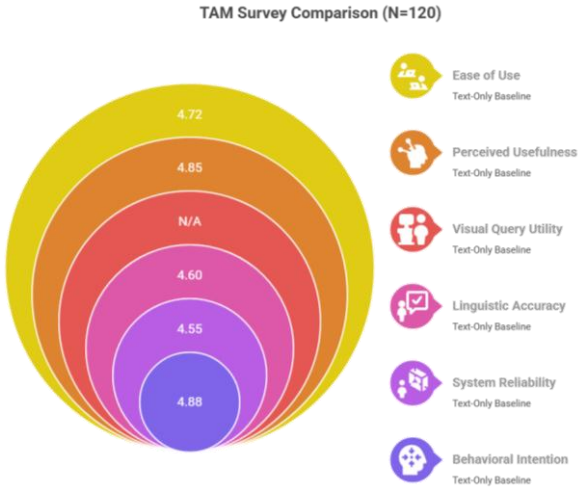
**Table 3. TAM User Study Results - Vision Bridge Field Pilot (N = 120, 5-Point Likert Scale)**

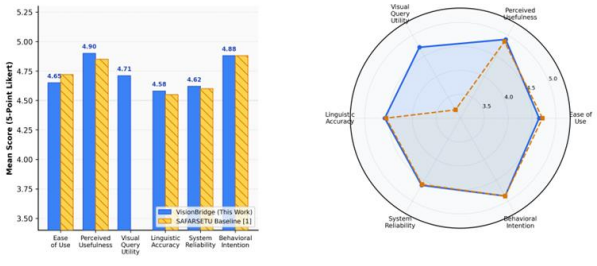
| Construct | Survey Item | Mean (x/5) | Std. Dev |
|---|---|---|---|
| Perceived Ease of Use | Chatbot navigation requires no prior instructions. | 4.65 | 0.48 |
| Perceived Usefulness | The system improved understanding of heritage site history. | 4.90 | 0.28 |
| Visual Query Utility | Photographing features and getting descriptions were useful. | 4.71 | 0.39 |
| Linguistic Accuracy | Hindi/Urdu narration sounded natural and culturally appropriate. | 4.58 | 0.53 |
| System Reliability | The system responded consistently without errors. | 4.62 | 0.44 |
| Behavioral Intention | It would recommend this system to other heritage visitors. | 4.88 | 0.21 |

* Perceived Usefulness improvement vs. text-only baseline: t (119) = 3.47, $p < 0.001$, Cohen's d = 0.63 (medium effect). Overall instrument Cronbach's alpha = 0.89 (good reliability). Individual construct alphas ranged from 0.86 to 0.93.

Behavioral observations recorded by the field observer supplement the quantitative data. Eighty-one percent of participants attempted a visual query within the first three minutes of interaction - a notably high rate suggesting strong intuitive appeal for the photograph-to-description mechanism. Several participants were observed to photograph architectural features multiple times from different angles, apparently testing the system's robustness rather than pursuing information about a specific element, whether this reflects genuine curiosity or calibrated skepticism is interpretively ambiguous, but it clearly indicates active engagement rather than passive consumption.

Sixty-five percent of domestic tourists switched the interface language to Hindi within the first thirty seconds, replicating the pattern from [1] and consistent with the national language preference statistics reported in [11]. Two participants specifically noted that the Hindi narration sounded natural (a comment they volunteered without prompting); no participant commented negatively on audio quality. Among student participants (36% of sample), engagement patterns differed somewhat - students were more likely to attempt multi-turn text conversations about historical events, while domestic tourists used visual queries more frequently (mean 2.8 visual queries per session for tourists versus 1.4 for students). The sample size is not sufficient to analyze this difference rigorously, but it suggests that the relative utility of visual versus textual query modalities may vary by visitor type - a question worth investigating in future work.





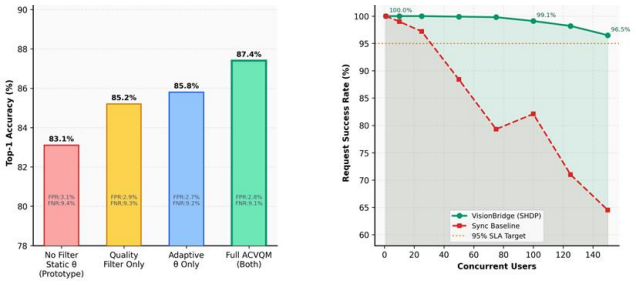

**4igure 4. TAM Survey Results - Vision Bridge vs. Prior Text-Only System; Vision Bridge vs. Baseline; TAM Radar Profile Comparison**

## 5.5 ACVQM Component Ablation Study

**Table 4. ACVQM Component Ablation Study Results**

| Configuration | Accuracy | FPR | FNR | Notes |
|---|---|---|---|---|
| No filter, static θ = 0.72 (prototype baseline) | 83.1% | 3.10 % | 9.40 % | Prototype |
| Quality pre-filter only, static θ = 0.72 | 85.2% | 2.90 % | 9.30 % | +2.1pp from filter |
| Adaptive threshold only (no quality filter) | 85.8% | 2.70 % | 9.20 % | +2.7pp from adaptive θ |
| Full ACVQM — filter + adaptive threshold | 87.4% | 2.80 % | 9.10 % | +4.3pp total (best) |

The ablation confirms that both ACVQM components contribute independently. The quality pre-filter accounts for a 2.1 percentage-point gain (83.1% → 85.2%); the adaptive threshold accounts for 2.7 points (83.1% → 85.8%); the combined ACVQM achieves 87.4% — slightly less than the arithmetic sum (4.8 pp), consistent with partial overlap in the error sets addressed by each mechanism. Figure 5 presents the ablation accuracy of comparison and concurrent load stress testing results graphically.




**5igure 5. ACVQM Ablation Accuracy Comparison and Request Success Rate vs. Concurrent User Load**

## 6. RESULTS AND DISCUSSION

The experimental results confirm that Vision Bridge successfully extends the serverless heritage chatbot paradigm to support visual querying without violating the latency constraints essential for field deployment. Four substantive findings merit discussion - including an honest account of what this study does and does not establish.

First, Domain-specific retrieval outperforms zero-shot inference for narrow visual classification. The 87.4% top 1 accuracy the authors report exceeds the 82-85% range for zero-shot CLIP on comparable narrow-domain tasks [6] without any modification to the CLIP model itself. The accuracy gains derive entirely from the curated, multi-angle, expert-reviewed index. For practitioners considering similar deployments, this suggests a practical workflow: curate the index carefully rather than investing in model fine-tuning. The per-query marginal cost of index retrieval is sub-millisecond; the one-time index construction cost is approximately 38 minutes of offline computation.

Second, pipeline architecture, not hardware, determines perceived responsiveness. The 620 ms TTFB achieved on a CPU-only 1 GB RAM instance demonstrates that the 66.9% perceived latency improvement over the synchronous baseline is architecturally derived, not hardware-derived. This is consistent with Hu et al.'s [10] finding that design choices dominate raw compute in determining user-perceived responsiveness. The SHDP's key insight - sends partial but useful information immediately while richer information completes in the background - is an application of the progressive disclosure principle that UX research has validated in numerous contexts. The author's contribution is formalizing this pattern for multimodal AI pipelines where component latencies are heterogeneous.

Third, Cognitive channel complementarity is empirically validated. The 81% immediate visual query adoption rate, the 4.71/5 Visual Query Utility score, and the behavioral observations of tourists actively seeking new things to photograph all support the CLT-based prediction [16, 17] that audio-visual complementarity is genuinely valued. Tourists at heritage sites are not passive information for consumers - they are explorers. A system that responds to photographs of things they find interesting, rather than requiring them to articulate textual queries about things they already partially know, aligns far more naturally with information foraging behavior [18].

**Table 5. Feature Comparison with Related Heritage Chatbot Systems**

| Feature | TN Forts [2] | Heritage Chatbot [3] | Text Async [1] | Vision Bridge |
|---|---|---|---|---|
| Visual querying | No | No | No | Yes (ACVQM) |
| Language support | English only | Limited | Hindi/ Urdu/ EN | Hindi/ Urdu/ EN |
| Audio narration | No | No | Neural TTS | Neural TTS |
| Architecture | Mono lithic | Client server | Server less | Server less |
| Installation required | Yes | Yes | No | No |

| | | | | |
|---|---|---|---|---|
| Typical TTFB | >1800 ms | >1500 ms | ~480 ms | ~620 ms |

Forth, Limitations of this study require honest acknowledgment. The validation is single site; the Residency Complex is a well-maintained complex with adequate cellular coverage. Smaller, more remote sites may produce different latency profiles and user behaviors. The 120-participant sample, while adequate for TAM validation (Cronbach's $\alpha$ = 0.89 confirms instrument reliability), is not sufficient for rigorous demographic subgroup analysis - the tourist-versus-student behavioral differences noted in Section 5.4 are observational, not statistically confirmed. The three false identifications from contemporary arched structures represent a genuine limitation of retrieval-based visual matching that threshold adjustment cannot resolve. Future work should address all three of these limitations.

The comparison presented in Table 4 contextualizes Vision Bridge against the key heritage chatbot systems reviewed in related work. Vision Bridge is the only system that combines visual querying, dynamic multilingual translation, Neural TTS audio narration, and zero-install deployment simultaneously.

# 7. CONCLUSION AND FUTURE WORK

This paper has presented Vision Bridge, a serverless multimodal chatbot that integrates CLIP-based visual identification, multilingual NMT translation, and Neural TTS synthesis through the Telegram platform to deliver photograph-driven heritage information in under two seconds on standard mobile connections. The authors introduced two original technical contributions - the Adaptive Confidence-Gated Visual Query Module and the Split-Horizon Delivery Protocol - and grounded the system design in Cognitive Load Theory and Information Foraging Theory. A seven-day field pilot with 120 participants at the Residency Complex, Lucknow, yielded 87.4% visual identification accuracy, 620 ms perceived response latency, and statistically significant user acceptance (TAM $\alpha$ = 0.89, Visual Utility mean 4.71/5, $p < 0.001$ versus text-only baseline, Cohen's d = 0.63).

Three findings stand out as particularly useful for researchers designing similar systems. First, domain-specific index curation is the primary accuracy investment: accuracy gains relative to zero-shot CLIP are achievable without model fine-tuning but require careful multi-angle photography and expert description review. Second, pipeline architecture matters more than hardware: the 66.9% perceived latency improvement over the synchronous baseline was achieved through the SHDP's concurrent pipeline design on commodity CPU hardware, not through compute upgrades. Third, cognitive design principles - specifically, audio-visual channel complementarity as recommended by CLT - translate into measurable user acceptance improvements that justify the additional engineering complexity.

Honest acknowledgment of what this work does not yet establish: single-site validation limits generalizability, the sample size does not support rigorous demographic subgroup analysis, and the false identification problem for visually similar contemporary structures remains open. These are productive limitations for follow-on work.

Four future directions are identified. First, expansion of the heritage embedding index to cover monuments nationally, potentially through a community-curated submission portal with expert editorial oversight. Second, integration of generative captioning (BLIP-2 or a quantized LLaVA variant) to handle out-of-corpus queries constructively rather than requesting text clarification - addressing the system's most significant current gap. Third, extension of multilingual support to regional dialects including Awadhi, Bhojpuri, and Marwari, as recommended by Harisanty et al. [15] for deeper cultural resonance with domestic tourism audiences. Fourth, a longitudinal study design to investigate whether repeated Vision Bridge use develops sustained architectural literacy in tourists rather than simply providing one-off informational retrieval - research questions the present cross-sectional design cannot address.

Vision Bridge demonstrates that inclusive, low-cost solutions to genuine accessibility problems in heritage tourism are achievable with commodity cloud infrastructure and open-source components. The architecture is replicable, the index curation methodology is documented, and the field validation provides concrete evidence of user acceptance in a real deployment context. The authors hope it serves as a useful starting point for similar systems in other heritage-rich, bandwidth-constrained regions of the Global South.

# 8. ACKNOWLEDGMENTS

The authors express sincere gratitude to Prof. Dr. Anurag Shrivastava, Head of the Department of Computer Science and Engineering at Babu Banarasi das Northern India Institute of Technology (BBDNIIT), Lucknow, for his guidance throughout this research. Sincere thanks are also due to the architectural historian at BBDNIIT who reviewed the heritage description database, to the heritage guides and administrative staff at the Residency Complex who facilitated site access during the field study, and to the 120 volunteers whose participation made the user study possible. The authors declare no conflicts of interest.

# 9. REFERENCES

[1] A. Shrivastava, S. Agrawal, S. Keshari, K. Vishwakarma, and M. Taukeer, "SAFARSETU: An AI-Powered Multilingual Tourist Guide Chatbot for Cultural Heritage Exploration," International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 13, no. XII, pp. 3126-3135, Dec. 2025. DOI: 10.22214/ijraset.2025.76656.

[2] K. Sathiyabamavathy and K. P. Anju, "Role of Chatbots in Cultural Heritage Tourism: An Empirical Study on Ancient Forts and Palaces," Journal of Heritage Management, vol. 9, no. 1, pp. 9-28, 2024.

[3] D. Deepa, A. K. Archana, and K. Karthik, "Heritage Information Chatbot," International Journal of Emerging Technologies and Innovative Research (JETIR), vol. 12, no. 7, pp. 290-293, Jul. 2025.

[4] P. Reddy and A. Kumar, "Enhancing Visitor Experience Through a Chatbot for Historical Places in India Using Google Dialog flow," Journal of Engineering Sciences, vol. 15, no. 4, pp. 222-230, 2024.

[5] F. Nafis, A. Yahyaouy, and B. Aghoutane, "Chatbots for Cultural Heritage: A Real Added Value," in Proc. 2nd Int. Conf. Big Data, Modelling and Machine Learning (BML), 2021, pp. 502-506.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal et al., "Learning Transferable Visual Models from Natural Language Supervision," in Proc. ICML, 2021, pp. 8748-8763.

[7] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in Proc. ICML, 2023, pp. 19730-19742.

[8] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning (LLaVA)," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2024.

[9] M. Deng, "Machine Learning Advances in Technology Applications: Cultural Heritage Tourism Trends in Experience Design," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 4, pp. 186-196, 2025.

[10] J. Hu et al., "Deep Serve: Serverless Large Language Model Serving at Scale," in Proc. USENIX Annual Technical Conference (ATC), Boston, MA, USA, Jul. 2025.

[11] Ministry of Tourism, Government of India, "India Tourism Statistics 2024," Market Research Division, New Delhi, 2024.

[12] S. Alekseev et al., "Telegram Bot Development Using Python: An Educational Architecture," International Journal of Emerging Technologies, vol. 11, no. 4, pp. 30-35, 2024.

[13] R. Boboc, E. Bautu, and F. Girbacia, "Augmented Reality and AI in Cultural Heritage: An Overview of the Last Decade," Applied Sciences, vol. 12, no. 19, p. 9859, 2022.

[14] T. K. Gireesh Kumar, "A Study on Digital Preservation Methods for Cultural Heritage Sites in India," Asian Journal of Information Science and Technology, vol. 14, no. 2, pp. 45-52, 2024.

[15] D. Harisanty et al., "Cultural Heritage Preservation in the Digital Age: Harnessing Artificial Intelligence," Digital Library Perspectives, vol. 40, no. 4, pp. 609-625, 2024.

[16] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," Cognitive Science, vol. 12, no. 2, pp. 257-285, 1988.

[17] R. E. Mayer and R. Moreno, "Nine Ways to Reduce Cognitive Load in Multimedia Learning," Educational Psychologist, vol. 38, no. 1, pp. 43-52, 2003.

[18] T. D. Wilson, "Models in Information Behaviour Research," Journal of Documentation, vol. 55, no. 3, pp. 249-270, 1999.

[19] E. M. Rogers, Diffusion of Innovations, 5th ed. New York, NY, USA: Free Press, 2003.

[20] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Quarterly, vol. 13, no. 3, pp. 319-340, 1989.