EROS INNOVATION

Culture: how it works.
Follow one scene through the benchmark.

Context & discovery

Cultural Atlas
Indian tourism references

Culture & heritage

Wikipedia · UNESCO
Sahapedia · IGNCA

Visual & regional evidence

Wikidata · Wikimedia Commons · ASI
Census of India · Ministry of Tribal Affairs

ACTUAL DATASET EXAMPLE · IG_G_001139

Magh Bihu dawn near Nagaon, Assam: a bamboo-and-hay meji bonfire, a bhelaghar hut beside it, and women offering til pitha and coconut laru.

01 · Input evidence

Start with the named sources above. Link each cultural claim to its specific supporting reference.

→

02 · Structured prompt

Claude Code helps turn the scene into a prompt and required visual checks.

→

03 · Machine verification

Validation gates + Gemma 4 31B check the claims against their evidence.

→

04 · Human approval

Experts review cultural accuracy and translations before test-set release.

→

05 · Model evaluation

Generate images from approved prompts and check each required detail. Planned.

Human check, in this example

Are the fire, hut and offerings correctly identified as Magh Bihu? Does the translation preserve those details? Machine verification alone does not release the scene.

Sources: repository reference registry + team-reported Indian tourism references. Example IG_G_001139 cites Magh Bihu; shortened prompt, human release pending.

EROS INNOVATION

Culture: what we have.
The data behind the example.

237
authored scenes

500
target scenes

10
categories

34
states / UTs

143 verified* 55 need experts 39 not yet verified

*Machine-verified. All 237 current scene records remain drafts.

Pending: expert checks, 1,356 translation rows and 263 more scenes to reach 500. Five-language design: EN / HI / TA / TE / BN.

ONE RECORD IS MORE THAN A PROMPT

Magh Bihu · Assam

Bihu dancers · Assam · Nayan j Nath · CC BY-SA 4.0 · display cropped

Core checks	Meji fire · bhelaghar hut · offerings
Supporting detail	Harvested field · regional clothing
Current state	Machine-verified → human review

Regional clothing reference; this photo does not depict the Magh Bihu ritual or a scored output.

Counts: current v0.3 canonical, verification and translation workspaces. Example: IG_G_001139. Translation rows include multiple prompt modes.

EROS INNOVATION

Culture: two kinds of expectation. Explicit details. Implicit meaning.

ILLUSTRATIVE MAGH BIHU PROMPT

"Magh Bihu in Assam: villagers gather at dawn around a bamboo-and-hay meji bonfire."

01 · EXPLICIT EXPECTATIONS

Did it show what we asked?

Details written directly in the prompt.

Example checks: villagers, dawn and the bamboo-and-hay meji bonfire.

VLM-assisted checks

A vision-language model inspects the image against a checklist. People verify mistakes and uncertain judgments.

02 · IMPLICIT EXPECTATIONS

Does the culture make sense?

Expectations inferred from the cultural context, even when unstated.

Example checks: does the ritual fit Magh Bihu? Are clothing and offerings appropriate to this Assamese scene?

Culturally knowledgeable people

Review meaning, regional context and stereotypes against evidence. Record valid local variations.

Paper's human rubric

0 no alignment · **0.5** partial · **1** complete

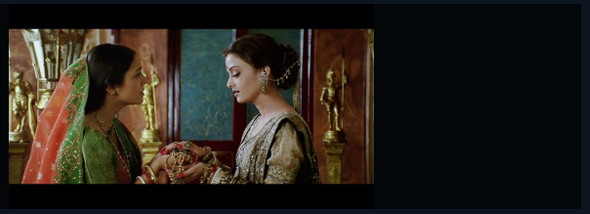
One alignment rating; below 1, flag **explicit / implicit / both** and explain why.

Our proposed workflow: VLM assistance + human review of both. CulturalFrames used humans for both; its automated metrics showed weak agreement with people. This is evaluation design, not a completed scoring run.

Reference: CulturalFrames (Nayak et al., 2025), §§3.3 & 6. Example adapted from IG_G_001139; not a paper prompt or scored output.

EROS INNOVATION

Cinema: the architecture.
Physical features → Gemma → people.



Devdas · real reference frame from the pilot

INPUT · REAL FILM FRAMES + MATCHED CROPS

Start with the craft inside the image

Example: compare a real frame with a copy whose grain has been removed. First measure the physical change; then use controlled image pairs to train and test the judge.

01 · PHYSICAL MEASUREMENT

Measure image features

Lane A measures 11 descriptors: grain, colour, lighting, depth of field and other craft cues.

Controlled changes check that each measurement reacts as intended.

02 · GEMMA FINE-TUNING

Train the AI judge

LoRA fine-tuning trials adapt Gemma 4 31B using real-versus-altered frame pairs and additional negative examples.

Gemma is the evaluator. The image-generating LoRA is a separate model.

03 · HUMAN EVALUATION

Validate the judgments

People compare images blindly, with model names hidden, and rate film look and content.

Check agreement with Gemma and test for bias before ranking generators.

In one line Measure the physical craft → teach Gemma what to look for → check it against people.

Completed: physical-feature pilot and Gemma fine-tuning trials. A reliable, human-validated generator ranking remains pending.

Sources: Lane A pilot, JUDGE_FINDINGS.md, JUDGE_STATUS_2026-09-06.md and current rating documentation. Fine-tuning trials do not imply acceptance.

EROS INNOVATION

Cinema: what we have.
A pilot sample and a larger training library.

69
films in pilot

6,900
pilot reference frames

550
Eros titles available*

1.3M
library frames*

Pilot generation: 150 prompts × 2 models × 2 seeds. Gemma: 9,600 scores; Muse: 231-score subset.

Pending: prepare the 550-title adaptation, reserve held-out films, validate the panel and schedule GPU capacity.

Bajirao Mastani · model-development reference.
Example visual material: lighting, costume and atmosphere.

Separate wider data engine*

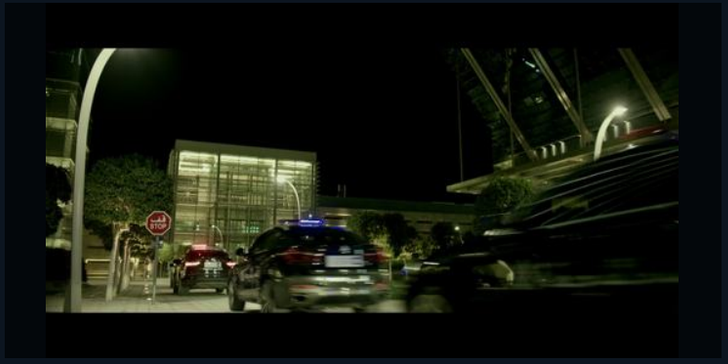
222.66M addressable → 8.29M processed → 6.43M captioned/exported.
3,008 GB; 77.6% yield.

Pilot: saved cinema records. *550-title / 1.3M-frame library and data-engine counts are team-reported snapshots, not live counters or completed full-library training.

EROS INNOVATION

Cinema: what gets judged?
One brief. Three images. Specific checks.

Shared brief: A wide night shot of a modern city street featuring a glass building, a police car with blue lights, and a blurred vehicle speeding past.



Real film reference
Dishoom

Base model output
Qwen-Image-2512 · seed 42

Eros LoRA output
Checkpoint 4250 · strength 0.5 · seed 42

Scene accuracy

Check the glass building, police car, blue lights and passing vehicle.

Cinematic craft

Compare night lighting, colour contrast, motion blur and wide-shot framing.

Human judgment

Rate blinded pairs. A selected example does not establish a winner.

Next run 1,000 held-out prompts × 8 models × 2 outputs **16,000 images**

Planned first wave: Qwen-Image-2512 base · current Eros LoRA · next Eros LoRA · i1-3B · FLUX.1 [Dev] · HiDream-I1-Full · Z-Image · Qwen-Image.

Dishoom pilot example p0060_8e2a9cd1f52e; both outputs use seed 42. Selected for illustration, not proof of a model winner.

EROS INNOVATION

Text benchmark.
Planned for a later phase.

We are currently testing different OCR models, including Surya and Sarvam. We will complete the Text benchmark later, after comparing the readers and settling the evaluation approach.

MODELS UNDER COMPARISON

Surya OCR · Sarvam · other OCR readers

Additional candidates include PaddleOCR / PaddleOCR-VL and Tesseract.

NOW

Test the readers

Run the same images through different models and compare how accurately they read Indian-language text.

NEXT

Review the results

Check errors, difficult scripts and unclear labels with human review. Choose a consistent evaluation setup.

LATER

Complete the benchmark

Finalise the Text benchmark and publish comparable results once the model evaluation is ready.

Current status Model comparison is in progress. The complete Text benchmark is deferred; no final model ranking is being presented.

Status: team update. Surya and Sarvam testing is ongoing; other reader candidates appear in repository configurations.