ChitraGati: A Grammar-Constrained Intelligent System For AI-Assisted Transcreation of Indian Folk Visual Art into Heritage Animation

Shrushti Kadam[1], Nikita Deore[2]
[1,2]*Department of Design, Analytical and Cyber Security, MIT Arts, Commerce & Science College, Alandi, Pune, India*
doi.org/10.64643/ijirt.208574-459

***Abstract*—Digitisation has served Indian folk art in one narrow sense. There are more photographs of Warli walls, Madhubani kohbars, Pattachitra pattas and Paithani pallus than at any earlier point in history. It has served far less well the part a practitioner would call essential: the rule system that decides what may be drawn, where it may sit, in which colour, and in what order. This paper treats that rule system as a first-class computational object. We define the Folk Visual Grammar Schema (FVGS), a typed and machine-readable encoding of a tradition's motif lexicon, admissible spatial syntax, chromatic constraints, compositional invariants and permissible motion ranges, and we use it not as training data but as an inference-time constraint on generative models. Around this schema we propose ChitraGati, a five-layer intelligent system for heritage animation. A perception layer performs folk motif recognition using a self-supervised vision transformer backbone with prototypical few-shot heads and open-set rejection, so that an unfamiliar motif is flagged rather than forced into the nearest known class. A generative layer performs AI-assisted animation generation: style-adapted latent diffusion with low-rank adapters and structural control produces assets, while a compact diffusion model over Motion Grammar Tokens, a discrete vocabulary of culturally attested kinetic primitives, produces motion in parameter space rather than pixel space, keeping output editable, auditable and stylistically bounded. A governance layer binds every generated asset to community-issued Traditional Knowledge Labels and cryptographically signed provenance manifests, and enforces a community-authored exclusion list for sacred imagery as a hard constraint rather than a soft prior. A pedagogy layer supplies an AI-based educational evaluation framework that traces learner mastery across a heritage concept graph, calibrates automatically generated assessment items, and closes the loop by feeding learning signals back to the generator so that segmentation, pacing and motif salience adapt to the learner. We introduce the Grammar Fidelity Score, a decomposable and computable measure of cultural fidelity, and the Artisan Concordance Index, which tests that measure against practitioner judgement instead of assuming agreement. The paper specifies the architecture, the guidance and loss formulations, a corpus and consent protocol, and a pre-registered evaluation plan with explicit success criteria. No empirical results are claimed at this stage; the contribution is a design, and specifically a design for letting generative systems move folk art without dissolving it.**

Index Terms*—Computer Vision, Generative AI, Diffusion Models, Neuro-Symbolic Systems, Motion Synthesis, Cultural Heritage Computing, Knowledge Tracing, Learning Analytics, Responsible AI, Indian Knowledge Systems (IKS).

I. INTRODUCTION

A Warli wall is not a picture of a village. It is an argument about how a village is arranged, made in a vocabulary so tightly bounded that a practitioner can tell within a second whether a figure belongs to the tradition or has been invented by someone who has only seen photographs of it. The same holds for Madhubani, where the choice between kachni line work and bharni colour fill changes what a panel is saying, and for Paithani, where the peacock in the pallu is not decoration but the resolution of a weave that has been building across the width of the fabric. These are rule systems. They constrain, and the constraint is the meaning.

Most digital heritage work does not preserve rule systems. It preserves surfaces. High-resolution scans, photogrammetry and museum catalogues capture how

a work looks and say almost nothing about what would have been permissible to draw next. For static archiving that omission is tolerable. For animation it is fatal, because animation is precisely the act of deciding what happens next, and a system with no model of permissibility will decide badly. Anyone who has watched a folk figure animated with default easing curves and a globally generic bounce has seen the failure mode: technically smooth, culturally illiterate.

The prior version of this research addressed that problem conceptually, proposing a transcreation framework in which folk visual grammars inform motion design for heritage learning. The framework was defensible but manual. It described what a sensitive designer ought to do; it did not describe a system that could do it, evaluate whether it had been done, or scale beyond the output of a single studio. This paper takes the same cultural commitment and rebuilds it as an intelligent system.

The timing is not incidental. Text-to-image and text-to-video models have become good enough that heritage institutions, schools and small studios are already using them on Indian folk material. They are also, on the current evidence, unreliable on exactly this material. Benchmarks of cultural competence in text-to-image models report that generations for under-represented cultural contexts collapse toward a small set of stereotyped renderings [22], and geographical audits find that outputs default to the visual conventions dominant in the training distribution rather than to the requested region [23]. A model asked for "Warli style" will typically produce a beige image with stick figures, which is close enough to satisfy a non-specialist and wrong in every respect a Warli artist would care about: the figures lose the two-triangle body, the chauk is absent or misplaced, and the composition has no ritual centre. The gap between plausible and correct is the gap this paper is about.

Our position is that the fix is not more training data, which does not exist in the required quantity and cannot ethically be manufactured by scraping artisan communities. The fix is to stop treating cultural style as a texture to be learned and start treating it as a grammar to be enforced. If the admissible motifs, relations, palettes and motions of a tradition can be written down in a form a machine can check, then a generative model does not need to have internalised the tradition. It needs only to be steered, at inference time, by something that has.

A. Problem Statement

We address four coupled problems. First, there is no machine-readable representation of Indian folk visual grammar, so cultural correctness cannot be computed and therefore cannot be optimised or audited. Second, existing generative pipelines produce folk-styled imagery whose fidelity is assessed by eye, informally and usually by non-practitioners. Third, motion for such imagery is authored either by hand, which does not scale, or by general-purpose video models, which drift stylistically across frames and offer no editable intermediate representation. Fourth, the educational claim routinely made for heritage animation, that it improves engagement and retention, is rarely instrumented; content is produced, published and never measured against learning.

B. Contributions

This paper makes six contributions.

C1. The Folk Visual Grammar Schema (FVGS): a formal, typed, machine-readable encoding of a folk tradition as a five-tuple of motif lexicon, spatial syntax, chromatic constraints, compositional invariants and kinetic ranges, aligned to established heritage ontologies and instantiated for Warli, Madhubani, Pattachitra and Paithani.

C2. A computer vision stack for folk motif recognition built for the data regime that actually exists: self-supervised pretraining on unlabelled corpora, prototypical few-shot heads for long-tail motifs, open-set rejection so that unknown motifs are surfaced rather than silently misclassified, and attribution maps that let an artisan audit the model's reasoning.

C3. Grammar-constrained generative synthesis, in which the perception stack is reused as a differentiable critic supplying a guidance gradient during diffusion sampling, so that grammar violations are suppressed during generation rather than filtered afterwards.

C4. Motion Grammar Tokens (MGT) and a compact conditional diffusion model that generates motion as interpretable parameters rather than pixels, together with a procedural-to-neural augmentation loop in which the symbolic schema synthesises its own labelled training data.

C5. The Grammar Fidelity Score (GFS), a decomposable cultural-fidelity metric over motif, chromatic, structural and kinetic components, and the Artisan Concordance Index (ACI), which measures how well GFS ranks agree with practitioner ranks instead of presuming that they do.

C6. An AI-based educational evaluation framework with a closed control loop, in which knowledge tracing over a heritage concept graph and privacy-preserving engagement telemetry adapt the generator's segmentation, pacing and salience decisions for the individual learner.

C. Scope and Honesty About Stage

This is an architecture and protocol paper. The system is specified in full, including model choices, objective functions, corpus design, consent procedure and evaluation plan, and every quantitative figure that appears in the tables is labelled as a design budget or a pre-registered success criterion. We have deliberately not reported measured results, because reporting them before the artisan partnerships described in Section VII are formally in place would violate the governance model the paper argues for. Section XI states plainly what these costs us in terms of validity.

D. Organisation

Section II reviews the relevant literature across five distinct areas and consolidates the gaps. Section III formalises FVGS. Section IV gives the system architecture. Sections V, VI, VII and VIII detail the perception, generative, governance and pedagogy layers respectively. Section IX presents the evaluation framework and success criteria. Section X analyses the expected effect on animation practice and economics. Section XI records limitations and an ethical risk register. Sections XII and XIII conclude and set out future work.

II. BACKGROUND AND RELATED WORK

A. Animation as a Medium for Heritage

The case for animation in heritage education rests on evidence from multimedia learning rather than on intuition. Mayer's cognitive theory of multimedia learning argues that learners process pictorial and verbal material through partially separate channels, and that gains come from managing the load on each: segmenting long sequences, signaling what matters, and keeping narration coherent with the image [11]. Meta-analytic work supports a real but conditional advantage for animation over static illustration, the condition being pedagogical alignment rather than production value [12], [13]. Traditional heritage documentation, however faithful, is static by construction and cannot exploit either the temporal channel or the narrative one [5], [6].

B. Folk Visual Grammars and the Temporality Already Inside Them

Indian folk traditions have documented internal structure: bounded motif sets, conventional composition, symbolic colour and narrative sequence [7], [8], [9]. Studies of cultural adaptation in animation confirm that this structure is what adaptation most often loses [10]. What is less often noted, and what matters for this work, is that several of these traditions already encode time.

Fig. 1. Warli motifs depicting daily life and communal activity. Figures are built from two triangles joined at the apex, and the dance chain spirals around a ritual centre rather than resting on a horizon line [1].

Fig. 2. Madhubani painting with fish motifs. The double-line border and the fully occupied field are syntactic constraints, not stylistic preferences [2].

Fig. 3. Pattachitra depicting mythological narrative. Episodes sit in bordered registers, which function as an authored panel sequence [3].

Fig. 4. Paithani pallu with peacock and asawali motifs. Because the form is woven, motif recurrence occurs at a fixed weave pitch, which behaves as a sampling rate [4].

Warli's tarpa dance is drawn as a spiral of joined figures around the musician. The composition is not a snapshot of a circle; it is a rotation rendered as accumulation, and it prescribes both the direction of travel and the fact that motion is cyclical rather than goal-directed. Bengal's closely related patachitra tradition goes further: the chitrakar unrolls a jorano pat vertically while singing the pater gaan, so that the scroll is a hand-driven timeline with an authored scrub rate and a fixed audio synchronisation. Odisha's Pattachitra, the tradition referenced in our earlier work, arranges episodes in bordered registers that function as a storyboard with explicit panel boundaries. Paithani, being woven, is discretised along the weft: motif repetition occurs at a fixed pitch, which is a sampling rate in everything but name, and the pallu operates as a cadence at the end of the fabric.

We treat these observations as evidence for the paper's central premise. Motion parameters for these traditions do not have to be invented by a designer or hallucinated by a model. To a useful extent they are recoverable from the source material, which means they can be written into a schema and enforced.

C. Computer Vision for Art and Motif Analysis

Vision transformers [14] provide the backbone architecture now standard for fine-grained visual recognition, and self-supervised pretraining, DINO [15] and DINOv2 [16] in particular, produces features that transfer well to domains with few labels. That property is decisive here: a folk-art corpus assembled ethically will have thousands of images, not millions. Few-shot recognition through prototypical networks [17] handles the long tail of rare motifs directly, by comparing an embedding to class centroids rather than by training a classifier per class. Detection and segmentation of individual motifs draw on DETR-style set prediction [18], Mask2Former [19] and promptable segmentation [20]. For a system that will be inspected by practitioners rather than by engineers, attribution matters as much as accuracy: Grad-CAM [21] and transformer-specific relevance propagation make it possible to show an artisan which strokes drove a classification.

Applied work on Indian folk art has so far concentrated on generation rather than analysis. Reported systems train GANs and, more recently, fine-tune diffusion models on Madhubani, Warli and related corpora, evaluating with SSIM, PSNR and Fréchet Inception Distance. These are reconstruction and distribution metrics. They quantify how close a generated image sits to a training distribution, which is not the same question as whether the image obeys the tradition's rules, and a model that has memorised the distribution can score well while producing a grammatically impossible composition.

D. Generative Models for Stylised Imagery

Denoising diffusion probabilistic models [24] and their latent-space formulation [25] are the current basis for high-resolution synthesis. Classifier-free guidance [26] provides the standard conditioning mechanism and, importantly for us, an insertion points for additional gradients. Personalisation to a small style corpus is well served by low-rank adaptation [27], DreamBooth [28] and textual inversion [29], all of

which can specialise a base model on the order of dozens to hundreds of images. Structural control through ControlNet [30] allows an external map, such as an edge or pose image, to constrain layout without retraining the base network. Earlier neural style transfer [31] and adversarial approaches [32] remain relevant as baselines.

The limitation common to all of these is that they encode style implicitly, in weights. There is no place in a LoRA adapter where one can read off the rule that a Warli chauk must be square, bounded and centrally placed, and therefore no place where that rule can be checked or corrected. Style mimicry is also the mechanism by which artists' work is appropriated at scale, which is why defensive tools such as Glaze [33] exist. A heritage system that participates in that mechanism, however well intentioned, is part of the problem.

E. Artificial Intelligence in Animation Production

Production-side automation is now credible for specific tasks. Learned inbetweening for hand-drawn content began with flow-based interpolation tuned to the discontinuities of animation [34] and correspondence-based colour and inbetween assistance [35], and has moved to generative interpolation: ToonCrafter adapts image-to-video diffusion priors to the cartoon domain and reports coherent interpolation between two drawings [36]. Motion itself can be generated directly. The Human Motion Diffusion Model [37] applies classifier-free diffusion to motion sequences and, notably for our design, predicts the clean sample rather than the noise so that geometric losses on joint positions and velocities remain usable. Temporal modules bolted onto image diffusion, such as AnimateDiff [38], and general video diffusion [39] produce plausible clips but expose no editable intermediate representation and drift in style over time. For heritage work that drift is not a cosmetic defect; a peacock that gradually stops being a Paithani peacock across sixty frames is a failure of the entire premise.

This is the reasoning behind our decision to generate motion in parameter space. A model that outputs amplitudes, periods, easing exponents and path curvatures produces something a designer can read, edit and version, something a grammar can bound with interval constraints, and something that renders deterministically. A model that outputs pixels produces something that must be inspected frame by frame and cannot be constrained except by rejection.

F. Artificial Intelligence in Educational Assessment

Learner modelling has a long lineage. Bayesian knowledge tracing [40] represents mastery of a skill as a latent binary variable updated by observed correctness, and remains a strong, interpretable baseline. Deep knowledge tracing [41] replaced the per-skill hidden Markov model with a recurrent network over interaction sequences and improved predictive accuracy at the cost of interpretability; attention-based successors [42] recovered part of that interpretability. Item response theory [43] supplies the calibration machinery needed to know whether an assessment item is discriminating or merely hard. Large language models now make it feasible to generate candidate items and to grade open responses against a rubric, though the reliability of such grading has to be reported, not assumed, and the standard practice of quoting agreement with human raters applies.

In heritage education specifically, evaluation is thin. Of the prior animation work surveyed in Table I, none instruments learning outcomes; effectiveness is asserted from viewing figures or from the enthusiasm of an audience. Normalised gain [44] and validated cognitive load instruments [45] are inexpensive to administer and would settle most of the open questions.

G. Responsible and Culturally Situated AI

Work on the politics of machine learning is directly applicable here. Decolonial critiques of AI [46] argue that systems built on extracted data reproduce the extraction; analyses of large-scale model behaviour warn about the flattening of minority representation [47]. Practical governance instruments exist and are under-used in technical papers. The CARE principles for Indigenous data governance [48] specify collective benefit, authority to control, responsibility and ethics as requirements on data about communities. Traditional Knowledge Labels issued through the Local Contexts initiative [49] let a community attach machine-readable permissions and protocols to a record. Content provenance standards, specifically C2PA Content Credentials [50], allow a cryptographically signed manifest that records what tool made an asset and from what. Model cards [51]

and datasheets [52] cover disclosure on the model and dataset side.

Our contention is that these are not appendices to a system like this one. They are components, and they can be enforced mechanically: a Traditional Knowledge Label that forbids ceremonial use can be checked before generation, not after publication.

H. Consolidated Research Gaps

Table I positions the present work against representative prior efforts. Five gaps recur.

G1. Cultural correctness is never formalised. It is assessed by eye, so it cannot be optimised, monitored, or contested with evidence.

G2. Generative work on Indian folk art evaluates with reconstruction and distribution metrics (SSIM, PSNR, FID) that are blind to rule violation.

G3. Motion is either hand-authored, and therefore unscalable, or produced by video models with no editable intermediate representation and measurable style drift.

G4. Learning outcomes are asserted rather than measured, and no reported system uses learner signals to change what is generated.

G5. Community consent, attribution and provenance are discussed as ethics sections rather than implemented as enforceable pipeline constraints.

TABLE I Comparison of Representative Prior Work with the Proposed System

Work / Year	Domain	AI Component	Grammar Formalised	Motion Handling	Cultural Metric	Learning Measured
Pune Mirror, 2023 [5]	Warli	None	No	Projection mapping, manual	Informal	No
AnimationXpress, 2023 [6]	Warli	None	No	Hand-animated short film	Informal	No
Mandal, 2025 [7]	Madhubani	None	Partial (motif catalogue)	Looping motifs, manual	None	No
Chathuranga, 2025 [8]	Folk arts	None	Partial (structural analysis)	Not implemented	None	No
OakLores, 2025 [9]	Pattachitra, Paithani	None	No	Storyboard adaptation	None	No
Folk-art GAN and diffusion studies	Madhubani, Warli	GAN, LoRA-tuned diffusion	No	Static images only	SSIM, PSNR, FID	No
ToonCrafter, 2024 [36]	General cartoon	Video diffusion	No	Generative interpolation	Perceptual only	No
MDM, 2023 [37]	Human motion	Motion diffusion	No	Parameter-space motion	Not applicable	No
ChitraGati (this work)	Four Indian folk traditions	CV, diffusion, LLM, knowledge tracing	Yes (FVGS)	MGT parameter-space diffusion	GFS validated by ACI	Yes, closed loop

III. THE FOLK VISUAL GRAMMAR SCHEMA

The schema is the part of this system that does the cultural work. Everything downstream, recognition, generation, evaluation, refers to it. We therefore define it precisely.

A. Definition

A folk visual grammar is represented as a five-tuple.

$G = (M, \Sigma, X, P, K)$

M is the motif lexicon: a finite set of typed symbols, each carrying an identifier, a semantic gloss, an iconographic role, an admissible-scale range and a sensitivity flag. Σ is the spatial syntax: a set of

production and adjacency constraints stating which motifs may co-occur, in which containment relations, and with what ordering along the reading direction of the tradition. X is the chromatic constraint set: admissible regions in CIELAB together with co-occurrence and ratio rules, since folk palettes are as much about proportion as about hue. P is the compositional invariant set: symmetry group, field-and-border partition, density bounds and the location of any ritual or narrative centre. K is the kinetic constraint set, mapping each motif class to intervals over motion parameters that the tradition's own imagery supports.

A distinguished subset $R \subseteq M$ marks restricted motifs. Membership in R is decided by the source community, not by the research team, and is enforced as a hard constraint at three points: corpus ingestion, generation-time masking, and pre-publication check. No amount of prompt engineering can lift it, because it is applied outside the model.

B. Encoding and Interoperability

FVGS instances are serialised as JSON-LD so that they can be published, versioned and diffed. Entity and event classes align to CIDOC CRM [53], the conceptual reference model used across museum informatics, and motif vocabulary is cross-referenced to the Getty Art and Architecture Thesaurus where a corresponding term exists. Alignment is not decoration. It means a heritage institution can ingest a schema instance without a bespoke importer, and it means the motion constraints travel with the cultural record instead of living in a studio's private project file.

C. Instantiation Across Four Traditions

Table II shows how the abstract tuple resolves for the four traditions in scope, and, in the final column, the kinetic primitives the source imagery supports. The mapping in that column is the operational heart of the transcreation argument: it is not an aesthetic preference but a claim about what the tradition's own structure licenses.

TABLE II Instantiation of the Folk Visual Grammar Schema Across Four Traditions

Tradition	Lexicon M (representative)	Syntax Σ and invariants P	Chromatic Set X	Licensed Kinetic Primitives K
Warli	Two-triangle human figure, tarpa player, chauk with Palghata, hut, tree, cattle, spiral dance chain	Chauk is square, bounded and centrally placed; figures orbit the musician; ground line is implicit not drawn	White rice-paste pigment on ochre or terracotta ground; effectively two-tone with restricted accent	LOOP-CYCLE (spiral orbit), STEP-PULSE (procession), FIELD-ACCRETION (figures appear in ritual order)
Madhubani	Fish, lotus, bamboo, parrot, sun and moon, kohbar composite, serpent	Double-line border mandatory; field is fully occupied; negative space is not idiomatic; kachni and bharni are not mixed within a panel	Saturated primaries and earth pigments; high-contrast outline against filled ground	AXIAL-SWAY (border oscillation), RADIAL-BLOOM (lotus opening), FILL-CASCADE (bharni hatching revealed progressively)
Pattachitra (Odisha)	Jagannath triad, Krishna Leela episodes, Dasavatara register, floral chita border	Episodes ordered in bordered registers; border is continuous and unbroken; frontal presentation of deities	Natural pigment palette, dominant red ground, restrained tonal range	PANEL-WIPE (register-to-register advance), REGISTER-HOLD (dwell on a narrative beat), BORDER-TRAVEL (chita as a progress indicator)
Paithani	Mor (peacock), asawali flowering vine, muniya parrot, bangdi mor, lotus, pallu cadence	Tapestry-weave repetition at fixed pitch; pallu terminates the composition; border and field are separately governed	Zari metallic thread against saturated silk ground; colour-shot (dhoop-chhaon) effect	WEFT-SHIMMER (metallic highlight travel), PITCH-REPEAT (motif recurrence at weave pitch), CADENCE-RESOLVE (pallu as narrative close)

Tradition	Lexicon M (representative)	Syntax Σ and invariants P	Chromatic Set X	Licensed Kinetic Primitives K
Bengal patachitra (reference)	Narrative scroll registers, chitrakar performance	Vertical scroll unrolled during sung pater gaan; register order is fixed by the song	Locally sourced natural pigment	SCROLL-ADVANCE (hand-driven timeline with authored scrub rate, audio-locked)

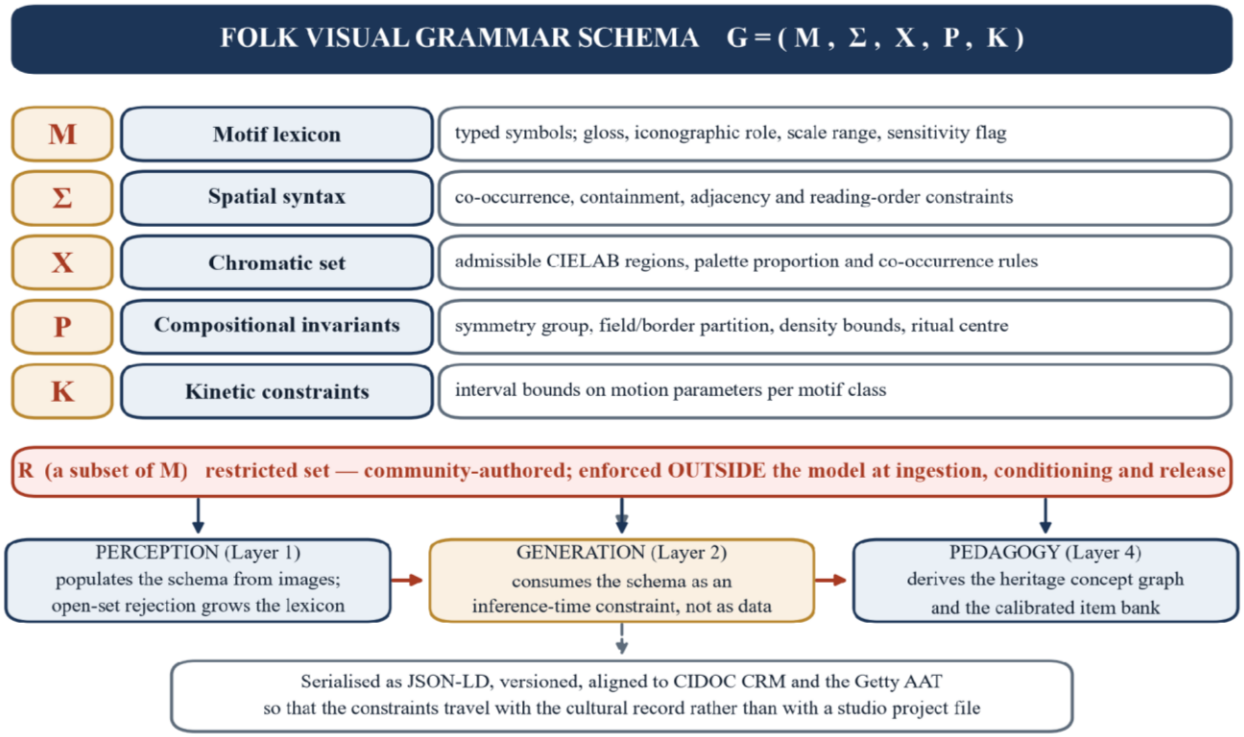

Fig. 5. The Folk Visual Grammar Schema and its downstream use. The five-tuple is populated by the perception layer, consumed as an inference-time constraint by the generative layer, and reused as the source of the heritage concept graph for assessment. The restricted set R is enforced outside the model at three checkpoints.

The final row is included because it supplies the clearest existing proof that these traditions carry authored timing. A jorano pat performance has a scrub rate, a synchronisation track and a fixed shot order. It is, functionally, an animation timeline that predates the medium by a considerable margin.

IV. SYSTEM ARCHITECTURE

ChitraGati is organised as five layers with a single shared artefact, the FVGS instance, passing between them. Fig. 6 shows the arrangement. The design rule we imposed on ourselves is that no layer may hold cultural knowledge implicitly in weights when it could hold it explicitly in the schema. Weights are for perception and plausibility; the schema is for permissibility.

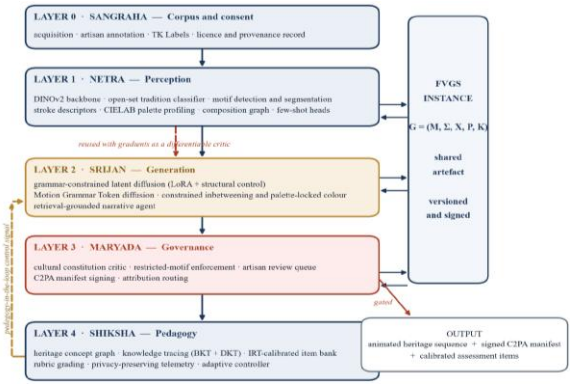

Fig. 6. ChitraGati system architecture. The Folk Visual Grammar Schema is the shared artefact; the perception layer both populates it and is reused as a critic that constrains the generative layer, while the pedagogy layer closes a control loop back onto generation.

Layer 0, Corpus and Consent (Sangraha). Acquisition, artisan annotation, Traditional Knowledge Labelling and licence recording. Nothing enters the system without a recorded provenance and permission state.

Layer 1, Perception (Netra). Computer vision for folk motif recognition, palette extraction, stroke characterisation and composition-graph construction. Populates FVGS instances from images and, later, scores generated candidates.

Layer 2, Generation (Srijan). Grammar-constrained asset synthesis, Motion Grammar Token diffusion for motion design, generative inbetweening and colourisation, and a retrieval-grounded language agent for narrative scripting.

Layer 3, Governance (Maryada). Cultural constitution checks, restricted-motif enforcement, artisan review queue, provenance manifest signing and attribution routing.

Layer 4, Pedagogy (Shiksha). Heritage concept graph, knowledge tracing, calibrated item bank, rubric grading and the adaptive controller that feeds learning signals back into Layer 2.

Two of these couplings are unusual and deserve to be named. The perception layer is not merely an analysis front end; it is reused, with gradients enabled, as the critic that steers generation. And the pedagogy layer is not a reporting dashboard; its output is a control input. Most systems in this space run analysis, generation and evaluation as three disconnected stages. Running them as a loop is what makes the system intelligent in the sense the term is meant to carry.

V. PERCEPTION LAYER: COMPUTER VISION FOR FOLK MOTIF RECOGNITION

A. The Data Regime, and Designing for It

An ethically assembled folk-art corpus is small. Our target specification, set out in Table V, is roughly 4,200 images across four traditions with an initial lexicon of about 180 motif types, annotated with artisan participation. Any architecture that needs a million labelled examples is therefore irrelevant regardless of its benchmark performance. The perception stack is designed around scarcity rather than apologising for it.

We pretrain a ViT-B/14 backbone with DINOv2-style self-distillation [16] on the unlabelled portion of the corpus together with a broader pool of Indian visual material, then keep the backbone frozen for most downstream heads. Self-supervision costs no labels and, in our judgement, is the single highest-leverage choice available in this domain.

B. Tradition Classification with Open-Set Rejection

A four-way classifier over traditions is trivially easy and trivially dangerous, because a closed-set classifier assigns every input to one of its classes, including inputs from traditions it has never seen. Gond, Cheriyal, Kalamkari and Sohrai would all be forced into the nearest of four boxes. We therefore reject rather than guess. Let f(x) be the backbone embedding and μ_c the prototype of class c. Class posteriors follows the prototypical formulation [17]:

$$p(y = c \mid x) = \exp(-d(f(x), \mu_c)) / \Sigma_{c'} \exp(-d(f(x), \mu_{c'}))$$

and the prediction is withheld when confidence or margin falls below calibrated thresholds:

predict c^* iff $\max_c p(y = c \mid x) \geq \tau$ and $d(f(x), \mu_{c^*}) \leq \delta$; otherwise UNKNOWN

Rejected inputs are routed to the artisan review queue, which is how the lexicon grows. An unknown motif is not an error to be suppressed; it is the most valuable thing the system can encounter, because it marks the boundary of what has been recorded.

C. Motif Detection, Segmentation and Stroke Characterisation

Individual motifs are localised with a set-prediction detector [18] and segmented with a mask transformer [19]; promptable segmentation [20] assists annotators during corpus construction and cuts labelling time substantially.

Stroke quality is characterised separately, because it is what distinguishes an authentic hand from a convincing imitation and it is not well captured by semantic features.

We compute a stroke descriptor combining stroke-width statistics, curvature distribution, junction-type histograms and responses from a bank of steerable filters at multiple orientations. In Madhubani this descriptor is what separates kachni from bharni; in Warli it captures the characteristic unmodulated line.

D. Chromatic Profiling

Palette extraction operates in CIELAB, where distances approximate perceptual difference better than in RGB. Each image yields a palette distribution obtained by weighted clustering with cluster counts fixed by the tradition's documented palette size. Chromatic conformity between a candidate image and the grammar's reference distribution X_{ref} is then the squared 2-Wasserstein distance:

$$E_{chroma}(x) = W_2^2(qLAB(x), X_{ref})$$

This choice matters practically. A histogram intersection would treat a small shift in every colour as catastrophic and a large shift in one colour as negligible; optimal transport gets the ordering right, which is what a colourist would also say.

E. Composition Graph Extraction

Detected motifs become nodes in a scene graph; edges encode containment, adjacency, alignment and reading order. This graph is the object against which spatial syntax is checked. Writing $\Gamma(x)$ for the composition graph of an image, structural conformity is a soft graph edit distance from that graph to the nearest admissible one:

$$E_{struct}(x) = \min_{g \in L(\Sigma, P)} \mathrm{GED}_{soft}(\Gamma(x), g)$$

where $L(\Sigma, P)$ is the language of admissible compositions. In practice the minimisation is approximated with a relaxed assignment, which is differentiable and therefore usable as a guidance signal.

F. Few-Shot Heads and the Long Tail

Motif frequency in any real corpus is heavily skewed. A handful of motifs appear constantly; most appear a few times. Prototypical heads over the frozen backbone let a new motif be added from five to ten annotated examples without retraining, which is the difference between a lexicon that grows with community input and one that is frozen at publication.

G. Explainability and the Artisan Audit Loop

Every classification surfaced to a human reviewer is accompanied by an attribution map, computed with Grad-CAM [21] for convolutional heads and relevance propagation for transformer heads. The purpose is specific and not decorative: an artisan reviewing a label needs to see whether the model attended to the diagnostic feature, the chauk boundary, the double border, the weave pitch, or to an incidental artefact of photography. Disagreements are logged with the attribution map attached, and the log is a training signal in its own right.

VI. GENERATIVE LAYER: AI-ASSISTED ANIMATION AND MOTION DESIGN

A. Grammar-Constrained Asset Synthesis

Asset generation uses a latent diffusion base [25] specialised per tradition with low-rank adapters [27] trained on the consented corpus, with structural conditioning [30] taking an edge or skeleton map derived from the FVGS instance rather than from an arbitrary reference image. This much is standard. The departure is what happens inside the sampling loop.

We define a grammar energy over a candidate image, assembled from the perception layer's components:

$$E_G(x) = w_m E_{motif}(x) + w_c E_{chroma}(x) + w_s E_{struct}(x)$$

and add its gradient to the classifier-free guidance update [26]. Writing $x_{0|t}$ for the clean sample predicted at step t, the modified noise estimate ε' is:

$$\varepsilon' = \varepsilon(x_t, t, \varnothing) + s\,[\,\varepsilon(x_t, t, c) - \varepsilon(x_t, t, \varnothing)\,] + \lambda\,\sigma_t\,\nabla_{x_t} E_G(x_{0|t})$$

The consequence is that grammar violations are suppressed while the image is still forming, when correction is cheap, instead of being detected after the fact and thrown away. Rejection sampling on a model that is wrong most of the time is expensive and, worse, biases the output toward whatever violations the checker happens to miss. Guidance biases the whole trajectory.

Restricted motifs in R are handled outside this mechanism. They are enforced by masking at the conditioning stage and re-checked before release, because a soft penalty on sacred imagery is not an acceptable design.

Algorithm 1: Grammar-Constrained Sampling

Input: prompt c, FVGS instance G, guidance scale s, grammar weight lambda, threshold tau_c

Output: asset x0 with GFS >= tau_c, or escalation to human review

1. x_T ~ N(0, I); attach G to the sampler context
2. for t = T down to 1 do
3. x0_hat <- denoise_estimate(x_t, t, c)
4. E <- w_m E_motif(x0_hat) + w_c E_chroma(x0_hat) + w_s E_struct(x0_hat)
5. eps_tilde <- CFG(x_t, t, c, s) + lambda sigma_t grad_{x_t} E

6. if any motif in R detected in x0_hat then re-mask conditioning and restart from t
7. x_{t-1} <- sampler_step(x_t, eps_tilde, t)
8. end for
9. compute GFS(x0); if GFS < tau_c then escalate to artisan review queue
10. return x0 with attribution maps and component scores

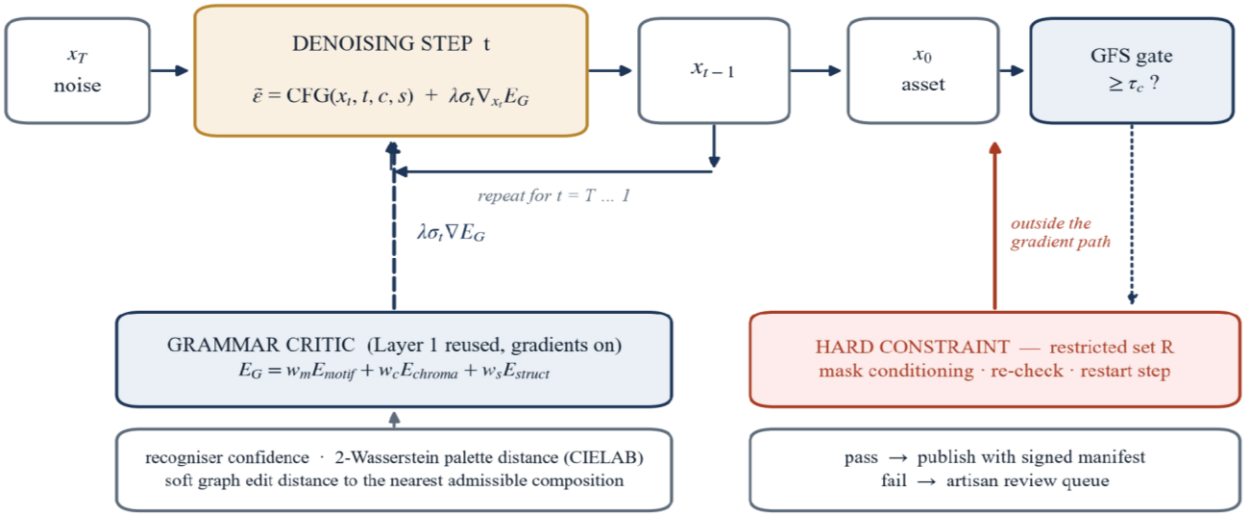

Fig. 7. Grammar-constrained diffusion sampling. The perception layer is reused with gradients enabled as a differentiable critic, so that motif, chromatic and structural violations are corrected during denoising rather than filtered after generation. Restricted motifs are handled outside the gradient path by hard masking.

B. Motion Grammar Tokens

Motion is where this system differs most from current practice. Rather than generating video frames, we generate the parameters of a motion, and we do so from a closed vocabulary.

A Motion Grammar Token is a named kinetic primitive attested in a tradition's own imagery, together with a parameter vector θ that instantiates it. The vocabulary used across the four traditions in scope is listed in Table III. Each token carries a parameter schema and, critically, interval bounds drawn from K in the FVGS instance.

TABLE III The Motion Grammar Token Vocabulary and Its Parameterisation

Token	Source Evidence	Parameters θ	Typical Bounds from K	Animation Use
LOOP-CYCLE	Warli tarpa spiral	angular velocity, radius, phase offset per figure, orbit eccentricity	period 2.0 to 6.0 s; eccentricity ≤ 0.25	Ritual and communal sequences
STEP-PULSE	Warli procession rows	step period, vertical amplitude, cohort phase lag	amplitude ≤ 4 per cent of figure height	Movement of groups across a field
AXIAL-SWAY	Madhubani border repetition	amplitude, period, easing exponent, axis angle	period 1.2 to 3.0 s; easing 1.5 to 2.5	Living borders and frames
RADIAL-BLOOM	Madhubani and Paithani lotus	scale rate, petal phase stagger, hold duration	stagger ≤ 120 ms per petal	Reveal of a symbolic centre
FILL-CASCADE	Madhubani bharni hatching	fill direction, stroke rate, ordering rule	stroke rate 20 to 60 per s	Progressive explanation of technique
PANEL-WIPE	Pattachitra registers	wipe direction, duration, border-lock flag	duration 0.4 to 1.2 s	Advance between narrative beats

Token	Source Evidence	Parameters θ	Typical Bounds from K	Animation Use
REGISTER-HOLD	Pattachitra dwell convention	hold duration, subtle-motion budget	hold 1.5 to 5.0 s	Comprehension pauses on a key event
BORDER-TRAVEL	Pattachitra chita continuity	travel rate, loop-closure constraint	must close within the shot	Progress indication without a UI element
WEFT-SHIMMER	Paithani zari dhoop-chhaon	highlight velocity, band width, angle	angle locked to weave direction	Material identity of silk and metal
PITCH-REPEAT	Paithani weave pitch	pitch, count, phase jitter	jitter ≤ 3 per cent of pitch	Rhythmic transitions and backgrounds
CADENCE-RESOLVE	Paithani pallu termination	deceleration profile, settle duration	settle 0.6 to 1.5 s	Scene and sequence endings
SCROLL-ADVANCE	Bengal jorano pat performance	scrub rate, audio lock points, register order	rate set by narration tempo	Long-form narration locked to audio

Motion is then generated by a compact conditional diffusion model over θ, conditioned on the motif's semantic class m, its narrative role r and the pedagogical intent pi supplied by Layer 4:

$\theta_0 \sim p_\varphi(\theta \mid m, r, \pi)$, subject to $\theta \in K(m)$

Following the motion diffusion literature [37], the network predicts the clean parameter vector rather than the noise, which lets us apply direct penalties on the parameters themselves. Training minimises reconstruction against artisan-approved exemplars plus a barrier term that grows sharply outside the admissible interval:

$L_{MGT} = E \| \theta_0 - \theta_0^{pred} \|^2 + \beta \Sigma_p \mathrm{ReLU}(|\theta_p - c_p| / h_p - 1)^2$

The model is small, on the order of a few million parameters over a six-layer transformer with model width 256, because the output space is a few dozen numbers rather than a few million pixels. It trains on a single consumer GPU in under an hour and samples in milliseconds. That efficiency is a design consequence, not a compromise: choosing the right representation made the model cheap.

C. Keyframe Assembly, Inbetweening and Colourisation

Sampled MGT parameters drive a deterministic rig that produces sparse keyframes from the generated assets. Inbetweens are produced by generative interpolation in the style of ToonCrafter [36], but constrained: the interpolator is conditioned on the MGT trajectory, so it fills in appearance rather than inventing motion. Colourisation is palette-locked to X, with any out-of-gamut pixel pulled back to the nearest admissible cluster in CIELAB. This division of labour, symbolic motion and neural appearance, is what prevents the style drift that pure video diffusion exhibits over long shots.

D. Retrieval-Grounded Narrative Scripting

A language model agent drafts shot lists and narration, which is the fastest part of production to accelerate and the most dangerous to leave ungrounded. Mythological and historical content is exactly where confident fabrication is most damaging and least likely to be caught by a general audience. We therefore restrict the agent to retrieval-augmented generation [54] over a curated, consented heritage corpus, and require every factual claim in a script to carry a citation span resolving to a retrieved passage. Claims that cannot be grounded are struck automatically and surfaced for human sourcing. The agent may write; it may not remember.

E. Grammar-in-the-Loop Data Augmentation

The schema is generative in its own right. Because Σ, X and P are explicit constraints, a procedural renderer can sample admissible compositions directly and emit them with perfect labels: motif identity, position, containment relation, palette assignment. These synthetic samples pretrain the perception heads, which then supervise the generative critic, which then produces better assets, which artisans annotate and fold back into the corpus. We call this procedural-to-neural bootstrapping, and it is the mechanism by

which a system with four thousand real images can behave as though it had considerably more without scraping a single additional artisan's work. The safeguard against synthetic collapse is a fixed cap on synthetic proportion per training batch and a held-out validation set that is real-only.

VII. GOVERNANCE LAYER: CONSENT, PROVENANCE AND CULTURAL SAFETY

Discussions of ethics in technical papers usually appear near the end and change nothing about the system. In ChitraGati the governance layer holds executable constraints, and generation fails closed when they are not satisfied.

A. Consent-First Corpus Construction

Every corpus item records the source community, the contributing artisan or institution, the licence, and one or more Traditional Knowledge Labels issued through the Local Contexts framework [49]. TK Labels are not licences; they are community-authored statements of protocol, and they map cleanly onto machine-checkable conditions. A TK Ceremonial label, for example, sets the corresponding motif class into R and blocks it from both training and generation. The CARE principles [48] govern the arrangement as a whole: collective benefit, authority to control, responsibility, ethics. A datasheet [52] accompanies the corpus and a model card [51] accompanies each released adapter.

B. The Cultural Constitution

Alongside the schema, partner communities' author a short natural-language document, the cultural constitution, stating what must not be done with their imagery. Examples from consultations of this kind are concrete: deities are not to be given comic expressions, ritual motifs are not to be used as decorative background, a marriage-chamber composition is not to be repurposed for commercial promotion. A language model critic evaluates each candidate asset and each script against this document in a critique-and-revise arrangement of the kind described for constitutional training [55]; failures are revised once and then escalated to human review. The critic is deliberately conservative, since the cost of a false positive is a designer's minute and the cost of a false negative is a community's trust.

C. Provenance and Attribution Routing

Every published asset carries a C2PA manifest [50] recording the base model, the adapter, the schema version, the corpus items retrieved, and the human reviewers who signed off. Two things follow. Downstream users can verify what they are looking at, which addresses the disclosure problem that arises whenever synthetic heritage imagery circulates. And because the manifest names the corpus items that influenced an asset, attribution and any revenue share can be routed to contributing artisans mechanically rather than by goodwill. We regard this as the minimum acceptable answer to the objection that generative models extract from the communities they imitate [46], [33].

D. Human Authority

Three decisions are reserved for people and cannot be delegated to the system: what enters R, whether a generated asset is culturally acceptable for release, and whether a disputed classification stands. The artisan review queue is not an appeals process bolted on at the end; items reach it by design, from open-set rejections, from low grammar-fidelity generations, from constitution failures and from a random audit sample drawn regardless of score.

VIII. PEDAGOGY LAYER: AN AI-BASED EDUCATIONAL EVALUATION FRAMEWORK

The educational claim made for heritage animation is testable and mostly untested. This layer makes it testable by construction, and then does something further: it uses the measurements to change the animation.

A. The Heritage Concept Graph

Assessment needs a target. We derive one directly from the schema: each motif, each syntactic rule and each chromatic convention becomes a node in a heritage concept graph, with prerequisite edges where understanding one concept presupposes another. Recognising a lotus precedes interpreting a kohbar; identifying a register precedes reading a Dasavatara sequence. Because the graph is generated from the FVGS instance, assessment coverage is guaranteed to match content coverage, which is a common failure point in hand-built curricula.

B. Knowledge Tracing Over Cultural Concepts

Learner mastery is tracked with a two-model arrangement. Bayesian knowledge tracing [40] provides an interpretable baseline whose four parameters (prior knowledge, transition rate, slip and guess) can be discussed with a teacher. Writing L_n for mastery of a concept after n opportunities, o_n for the observed response and P(T) for the transition rate:

$$P(Ln) = P(Ln_{-1} \mid o_n) + [\,1 - P(Ln_{-1} \mid o_n)] \cdot P(T)$$

A deep model [41], [42] runs alongside it over the interaction sequence:

$$h_t = \text{LSTM}(xt, h_{t-1});\ y_t = \sigma(W h_t + b)$$

We report both. The deep model predicts better; the Bayesian model explains better; a teacher needs the second even when the system acts on the first.

C. Automatically Generated, IRT-Calibrated Items

Assessment items are drafted by a language model constrained to the concept graph and to retrieved corpus passages, then filtered by an expert panel, then calibrated with a two-parameter item response model [43]:

$$P(uij = 1 \mid \zeta_i) = 1 / (1 + \exp[-a_j(\zeta_i - b_j)])$$

Here ζ_i is learner ability, a_j item discrimination and b_j item difficulty; ζ is used rather than the conventional θ because θ already denotes motion parameters in Section VI. Items with discrimination a_j below 0.4 are retired automatically. Generation makes a large item bank affordable; calibration is what makes it trustworthy. Neither step is sufficient alone, and we would regard an uncalibrated LLM-generated item bank as a liability rather than a feature.

D. Grading Open Responses

Free-text answers, for instance "explain why the fish appears in this panel", are graded against a rubric by a language model with retrieval grounding. Reliability is reported rather than assumed: a stratified ten per cent sample is double-marked by human raters and quadratic weighted kappa is published with every deployment. If agreement falls below the pre-registered floor, automated grading is suspended for that item type. This is ordinary measurement practice and it is routinely skipped.

E. Engagement Without Surveillance

Engagement is estimated from interaction telemetry only: dwell time per segment, replay and scrub-back events, pause location relative to annotations, and response latency. We take no webcam, face, gaze or affect data. The refusal is deliberate. Emotion recognition on classroom video is of contested validity and unambiguously invasive, and the marginal signal it might add does not justify the cost to learners who cannot meaningfully consent. Telemetry alone is sufficient for the control decisions the system actually makes.

F. Pedagogy-in-the-Loop Generation

Here the loop closes. The controller selects presentation actions, segment granularity, narration tempo, motif salience, choice of MGT for an emphasis beat, and the amount of REGISTER-HOLD dwell, to maximise expected learning gain, where Δm is the change in traced mastery, e is engagement and ℓ is reported cognitive load, subject to a hard cultural constraint:

$$\max E[\,\alpha \Delta m + \beta e - \gamma \ell] \text{ subject to GFS} \geq \tau_c$$

Formulated as a contextual bandit over the learner's traced state, the controller is directly operationalising Mayer's segmenting, signalling and coherence principles [11] instead of applying them as fixed authoring advice.

The constraint is what keeps the arrangement honest. Without it, the controller would eventually discover that exaggerated motion and saturated colour raise short-term engagement, and it would degrade the tradition to do so. The constraint GFS $\geq$ τc makes cultural fidelity non-negotiable and lets optimisation operate only in the space of choices the grammar already permits.

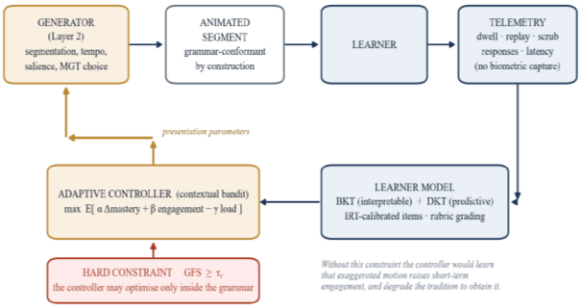

Fig. 8. The pedagogy-in-the-loop control arrangement. Learner signals adjust presentation parameters, but the cultural-fidelity constraint bounds the action space so that engagement can never be purchased at the cost of grammar.

IX. EVALUATION FRAMEWORK AND PRE-REGISTERED CRITERIA

A. The Grammar Fidelity Score

GFS aggregates the perception layer's components into a single interpretable score in the unit interval:

$$GFS(x) = w_1 M(x) + w_2 [1 - C(x)] + w_3 [1 - S(x)] + w_4 [1 - K(x)]$$

M is mean motif-recognition confidence with an open-set margin penalty, C is normalised chromatic divergence, S is normalised structural deviation, and K is kinetic violation computed over the realised motion parameters, with c_p and h_p the centre and half-width of the admissible interval for parameter p:

$$K = (1 / |P|) \Sigma_p \text{ ReLU } (| \theta_p - c_p | / h_p - 1)$$

Weights are elicited from practitioner panels per tradition rather than set by the authors, and are published with each schema version. Reporting the components separately is as important as reporting the total, because a single number hides which axis failed and an animator needs to know whether the problem is palette or composition.

B. The Artisan Concordance Index

A metric of cultural fidelity that has not been checked against culture bearers is an assertion. ACI is the Spearman rank correlation between GFS ordering and the ordering produced by a blinded practitioner panel over the same candidate set, with panel internal agreement reported as Krippendorff's alpha. If ACI is low, GFS is wrong and must be reweighted or redesigned; the metric is subordinate to the judgement, not a substitute for it.

C. Metric Suite

A system that makes four separate kinds of claim needs four separate kinds of evidence. Table IV sets out the full suite, deliberately including axes that are inconvenient for us: governance leakage, production cost and cognitive load are all places where a proposal of this sort can look good on paper and fail in a studio or a classroom.

TABLE IV The Evaluation Metric Suite

Dimension	Metric	Instrument	Purpose
Cultural fidelity	GFS and its four components	Perception layer	Machine-checkable grammar conformance
Cultural validity	ACI (Spearman rho), Krippendorff alpha	Blinded artisan panel	Tests whether GFS tracks expert judgement
Perception quality	Macro-F1, open-set AUROC, few-shot 5-shot accuracy	Held-out real-only test split	Recognition reliability, especially on the tail
Generative quality	FID, KID, CLIPScore, LPIPS	Standard implementations	Comparability with prior generative work
Temporal quality	FVD, inter-frame flicker index, MGT bound-violation rate	Rendered sequences	Detects style drift and out-of-grammar motion
Learning outcome	Normalised gain, two-week retention, QWK for auto-grading	Pre-test, post-test, delayed test	Whether the animation teaches
Cognitive load	Validated multi-component load scale	Post-session instrument	Distinguishes germane from extraneous load
Production efficiency	Artist-hours per finished second, revision count, cost	Production logs	Whether the system is worth adopting
Governance	Restricted-motif leakage rate, manifest completeness, review latency	Audit logs	Whether the safeguards actually hold

D. Planned Study Design

Validation proceeds in three studies. Study A is a technical benchmark of the perception stack on a real-only held-out split with ablations removing self-supervised pretraining, few-shot heads and open-set rejection. Study B is a generation study comparing four conditions, base diffusion with a style prompt, LoRA-adapted diffusion, LoRA with post-hoc filtering, and full grammar-constrained sampling, scored by GFS and by a blinded artisan panel of at least nine practitioners across the four traditions. Study C is a learning study with school and undergraduate cohorts on Maratha-era heritage content of the kind presented at Shivsrushti,

randomising static illustration, conventional animation, and ChitraGati output with and without the adaptive loop, measuring normalised gain, two-week retention and cognitive load.

E. Pre-Registered Success Criteria

The following are targets fixed in advance, not measurements. Stating them before data collection is what makes the eventual result interpretable.

TABLE V Pre-Registered Success Criteria and Design Budgets

Item	Target	Rationale
Corpus size and lexicon	About 4,200 consented images; about 180 motif types; 4 traditions	Achievable with partner institutions within one year
Artisan annotators	At least 12, at least 2 per tradition	Enough for inter-annotator agreement per tradition
Tradition classification macro-F1	≥ 0.90 on real-only held-out data	Four visually distinct classes; below this the pipeline is unsound
Open-set AUROC on unseen traditions	≥ 0.85	Must reliably flag Gond, Cheriyal and Kalamkari as unknown
Few-shot motif accuracy at 5 shots	≥ 0.75	Determines whether the lexicon can grow from community input
GFS improvement, constrained over unconstrained	≥ 0.15 absolute	The core claim of the generative layer
Artisan Concordance Index	$\rho \geq 0.70$ with panel $\alpha \geq 0.67$	Below this, GFS is not measuring what it claims
Restricted-motif leakage	0 occurrences in audit; any occurrence halts release	Non-negotiable safety property
MGT bound-violation rate	< 1 per cent of sampled parameters	Motion must stay inside the grammar
Normalised learning gain over static baseline	≥ 0.30 (medium effect)	Standard threshold in multimedia learning studies
Auto-grading agreement	QWK ≥ 0.70 against human raters	Below this, automated grading is suspended
Compute budget	Single 12 GB consumer GPU; adapter training under 4 hours per tradition	Deployability in an Indian college or small studio
Inference budget	Under 5 s per 512 × 512 keyframe; MGT sampling under 50 ms	Interactive authoring rather than batch rendering

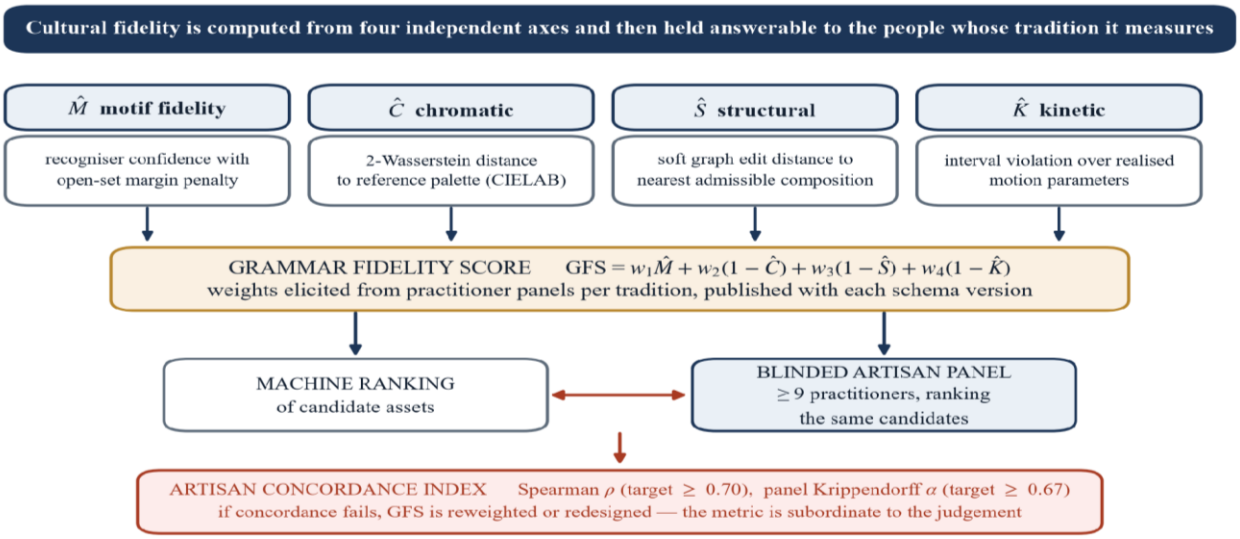

Fig. 9. Decomposition of the Grammar Fidelity Score and its validation path. The score is computed from four perception components; the Artisan Concordance Index tests the score itself against blinded practitioner ranking, so the metric remains subordinate to human judgement.

X. EXPECTED IMPACT ON ANIMATION PRACTICE

A. Where the Time Actually Goes

Traditional two-dimensional animation of culturally specific material is expensive in a particular way. The costly stages are not the creative ones. Reference gathering, style-sheet construction, asset variation, inbetweening, clean-up and colour flatting consume the majority of the schedule, and they are the stages where a folk tradition is most often compromised, because a studio under deadline will approximate a motif rather than research it. Table VI gives our design estimate for a two-minute educational sequence. These are budget figures used to size the system, not measured outcomes, and Study C is the study that would confirm or refute them.

TABLE VI Design Estimate for a Two-Minute Heritage Sequence

Production stage	Conventional practice	With ChitraGati	Nature of the change
Reference and style research	3 to 5 days	1 day	Schema and corpus already encode the tradition
Style sheet and asset variation	5 to 8 days	1 to 2 days	Grammar-constrained synthesis with artisan review
Motion design and keyframing	6 to 10 days	2 to 3 days	MGT sampling proposes, animator selects and edits
Inbetweening and clean-up	8 to 12 days	2 to 4 days	Generative interpolation constrained by MGT trajectories
Colour flatting	3 to 4 days	Under 1 day	Palette-locked colourisation
Cultural review	Ad hoc or absent	Continuous, logged	Review becomes a pipeline stage with an audit trail
Learning instrumentation	Rare	Built in	Assessment derived from the same schema

The interesting number is not the total. It is the redistribution. Cultural review moves from an optional courtesy to a logged pipeline stage, and instrumentation moves from rare to automatic, precisely because the schema that constrains generation is also the schema that generates the assessment.

B. What Changes for Whom

For small studios and college departments, the effect is access. A four-person team on consumer hardware can attempt culturally specific work that currently requires a studio with a research budget. For heritage institutions, museums, state tourism boards, sites such as Shivsrushti, the effect is that content can be produced in the volume that interpretation genuinely needs, in regional languages, without each new sequence being a fresh negotiation with cultural accuracy. For educators, the effect is that animation stops being a fixed artefact and becomes an adaptive one.

For artisans, the effect is the one we care most about and the one most easily got wrong. The default trajectory of generative AI in this space is extraction: scrape the images, learn the style, remove the artist. ChitraGati is arranged to invert the incentive. Artisans define the lexicon, decide what is restricted, hold veto over releases, are named in signed provenance manifests, and can be paid through attribution routing derived from those manifests. Whether this happens depends on institutional will rather than on architecture, and we make no claim that the architecture is sufficient. It is, however, necessary, and it is absent from every comparable system we surveyed.

C. What Does Not Change

The system does not decide what is worth telling. It does not adjudicate cultural meaning, and where the grammar is contested, and grammars are contested, in Warli particularly, where commercial adaptation has been disputed for decades, it records the disagreement rather than resolving it.

It proposes; people dispose. An animator who accepts every MGT sample without editing will produce competent, forgettable work, which is roughly what an animator who accepts every default easing curve produces today.

XI. LIMITATIONS, THREATS TO VALIDITY AND ETHICAL RISKS

We list these plainly, since a paper that proposes a governance layer should be willing to apply the same standard to itself. Table VII summarises the risks that survive the mitigations described above, together with the exposure that remains.

No empirical validation is reported. The architecture is specified and internally coherent; it has not been built and measured, and every quantitative figure in this paper is a target. Readers should treat the contribution as a design and a protocol.

Formalisation can itself be a form of flattening. Writing a grammar down fixes it, and living traditions change; a schema that is treated as authoritative could ossify practice or privilege one lineage's conventions over another's. Our mitigations are versioning, explicit provenance for every rule including which community authored it, and support for multiple coexisting schema variants per tradition. We do not claim these are sufficient, and we think this is the deepest unresolved risk in the work.

Coverage is narrow. Four traditions, one of which enters only as reference evidence, out of a folk landscape numbering in the dozens. Generalisation to traditions with different structural logic, Gond's dot-and-dash fill, Kalamkari's pen-work narrative density, is asserted rather than demonstrated.

Synthetic augmentation risks model collapse if the synthetic proportion is allowed to grow. We cap it and hold out a real-only validation set, which limits but does not eliminate the risk.

Language model components can fabricate. Retrieval grounding with mandatory citation spans reduces this materially but does not remove it, and mythological content is exactly where a plausible fabrication is least likely to be challenged by a general audience.

The learning studies face the usual threats: novelty effects inflating short-term engagement, cohort effects, and the difficulty of blinding participants to an obviously different visual treatment. Delayed retention testing and an active control, conventional animation rather than static illustration alone, address the first and third partially.

TABLE VII Ethical Risk Register and Mitigations

Risk	Severity	Mitigation in the architecture	Residual exposure
Sacred or restricted imagery generated	Critical	Set R enforced at ingestion, conditioning and release; audit sampling	Depends on completeness of R, which communities' control
Style extraction without benefit to artisans	High	Consent-first corpus, TK Labels, signed manifests, attribution routing	Enforcement is institutional, not technical
Cultural flattening by a fixed schema	High	Versioning, per-rule provenance, coexisting variants	Structural; not fully solvable by design
Fabricated mythology or history in narration	High	RAG with mandatory citation spans; ungrounded claims struck	Retrieval corpus gaps
Learner surveillance	Medium	Telemetry only; no biometric, gaze or affect capture	Telemetry is still behavioural data and needs consent
Automated grading error	Medium	IRT calibration, QWK floor, automatic suspension	Rubric design quality
Optimisation degrading fidelity for engagement	Medium	Hard GFS constraint on the controller action space	Threshold setting is a judgement call
Synthetic data collapse	Low to medium	Proportion cap, real-only validation	Long-horizon drift

XII. CONCLUSION

The earlier form of this research argued that Indian folk visual grammars could inform motion design for heritage learning, and demonstrated it conceptually. This paper argues something stronger and more specific: that those grammars can be written down in a form a machine can check, and that writing them down is what allows contemporary generative models to be used on this material without damaging it.

The Folk Visual Grammar Schema turns cultural correctness into something computable. The perception layer populates and then polices it, with open-set rejection ensuring that the system knows the

limits of its own lexicon. The generative layer uses it as an inference-time constraint rather than as training data, and generates motion as bounded, editable, interpretable parameters through the Motion Grammar Token vocabulary rather than as pixels. The governance layer makes consent, restriction and attribution executable rather than aspirational. The pedagogy layer measures whether any of it teaches, and returns that measurement to the generator as a control signal. The Grammar Fidelity Score makes fidelity reportable; the Artisan Concordance Index keeps that score answerable to the people whose tradition it claims to measure.

The wider argument is about direction of travel. Generative AI applied to indigenous art currently defaults to imitation: learn the surface, reproduce the surface, and let the community find out afterwards. The alternative proposed here is constraint: encode what the tradition permits, hand that encoding authority over the model, and let the community hold the pen. That is a harder system to build and a slower one to deploy. We think it is the only version worth building, and we have specified it in enough detail that it can be built, tested and, if it fails, shown to have failed.

XIII. FUTURE WORK

The immediate work is construction and the partnerships that must precede it. Corpus assembly with Warli artists in Palghar, Mithila painters, Raghurajpur chitrakars and Yeola weavers has to be arranged under CARE terms before any model is trained, and the schema for each tradition has to be authored with, not merely reviewed by, those practitioners. Studies A, B and C then follow in order. Beyond that, four directions look productive. Extending FVGS to traditions with different structural logic, Gond, Cheriyal, Kalamkari, Sohrai, would test whether the five-tuple is general or has quietly been fitted to four cases. Learning kinetic constraints directly from recorded performance, motion capture of tarpa dance, high-frame-rate video of a patua unrolling a scroll, would replace elicited intervals with measured ones and is the most obvious way to strengthen K. Extending the output beyond flat animation to real-time engines, augmented reality overlays at heritage sites and volumetric exhibits would make the grammar constraints operate under interactive rather than authored timing. And the governance instruments deserve study in their own right: whether attribution routing derived from provenance manifests actually delivers income to artisans is an empirical question about institutions, and it should be answered with the same rigour we are asking of the technical claims.

Finally, the schema itself should be released. A published, versioned, community-owned FVGS for even one tradition would be more durable than any model trained on it, since models are superseded every few months and a well-formed grammar is not.

REFERENCES

[1] “Warli Art,” Chukkimane. [Online]. Available: Accessed: Sep. 7, 2025.
[2] “Traditional Madhubani Painting, Folk Art, Fish Art, Indian Art,” Etsy. [Online]. Available: Accessed: Sep. 7, 2025.
[3] “Lord Jagannath, Subhadra, and Baladev Pattachitra Painting,” Etsy. [Online]. Available: Accessed: Sep. 7, 2025.
[4] “Asawali Brocade Border Paithani Collection,” Touch of Class Paithani. [Online]. Available: Accessed: Sep. 7, 2025.
[5] “When Walls Dance: A Captivating Fusion of Dance and Warli Art,” Pune Mirror, Nov. 2023. [Online]. Available:
[6] “Nature Conservation, Looking Beyond Virtual Life: Ekabhuya Animation Raises Strong Voice,” AnimationXpress, 2023. [Online]. Available:
[7] A. Mandal, “Development of Motifs from Madhubani Painting: A Study in Bihar,” ResearchGate, 2025. [Online]. Available:
[8] E. M. Chathuranga Ekanayaka, “An Analytical Study on the Influence and Reinterpretation of Indian Folk Art in Contemporary Artistic Expression,” Journal of Novel Research and Innovative Development, vol. 3, no. 5, pp. 1–6, May 2025.
[9] “Pattachitra Painting: Odisha's Timeless Saga of Art and Mythology,” OakLores, Mar. 2025. [Online]. Available:
[10] “An Initial Attempt: A Synthesis of Cultural Adaptation and Representation in Animation,” ResearchGate, 2021. [Online]. Available:
[11] R. E. Mayer, Multimedia Learning, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[12] T. N. Hoffler and D. Leutner, "Instructional animation versus static pictures: A meta-analysis," Learning and Instruction, vol. 17, no. 6, pp. 722–738, 2007.
[13] S. Berney and M. Betrancourt, "Does animation enhance learning? A meta-analysis," Computers and Education, vol. 101, pp. 150–167, 2016.
[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
[15] M. Caron et al., "Emerging properties in self-supervised vision transformers," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 9650–9660.
[16] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," Transactions on Machine Learning Research, 2024.
[17] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Proc. European Conf. Computer Vision (ECCV), 2020, pp. 213–229.
[19] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1290–1299.
[20] A. Kirillov et al., "Segment anything," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023, pp. 4015–4026.
[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626.
[22] N. Kannen et al., "Beyond aesthetics: Cultural competence in text-to-image models," in Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2024.
[23] A. Basu, R. Venkatesh Babu, D. Pruthi et al., "Inspecting the geographical representativeness of images from text-to-image models," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023.
[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 6840–6851.
[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and P. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
[26] J. Ho and T. Salimans, "Classifier-free diffusion guidance," arXiv preprint arXiv:2207.12598, 2022.
[27] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in Proc. Int. Conf. Learning Representations (ICLR), 2022.
[28] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22500–22510.
[29] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," in Proc. Int. Conf. Learning Representations (ICLR), 2023.
[30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023, pp. 3836–3847.
[31] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414–2423.
[32] I. J. Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems (NeurIPS), vol. 27, 2014.
[33] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," in Proc. 32nd USENIX Security Symposium, 2023.
[34] L. Siyao et al., "Deep animation video interpolation in the wild," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6587–6595.

[35] E. Casey, V. Perez, and Z. Li, "The animation transformer: Visual correspondence via segment matching," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 11323–11332.

[36] S. Devikar, M. Hinge, and S. S. Shevkari, "Human vs. machine: A review of AI's impact on language learning interaction or replacing human interaction," International Journal for Multidisciplinary Research, vol. 7, no. 3, 2025, doi: 10.36948/ijfmr. 2025.v07i03.49459.

[37] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," in Proc. Int. Conf. Learning Representations (ICLR), 2023.

[38] Y. Guo et al., "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning," in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[39] A. Blattmann et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," arXiv preprint arXiv:2311.15127, 2023.

[40] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," User Modeling and User-Adapted Interaction, vol. 4, no. 4, pp. 253–278, 1994.

[41] C. Piech et al., "Deep knowledge tracing," in Advances in Neural Information Processing Systems (NeurIPS), vol. 28, 2015, pp. 505–513.

[42] A. Ghosh, N. Heffernan, and A. S. Lan, "Context-aware attentive knowledge tracing," in Proc. 26th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2020, pp. 2330–2339.

[43] F. M. Lord, Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ, USA: Lawrence Erlbaum, 1980.

[44] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," American Journal of Physics, vol. 66, no. 1, pp. 64–74, 1998.

[45] H. Chinni, Y. K. Sharma, S. Shevkari, J. L. Vydehi, Saurabh, and M. Kavitha, "Comparative Evaluation of Custom CNN Architectures and Transfer Learning Models for Image Classification in PyTorch," in Proc. 2026 13th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2026, pp. 1–6, doi: 10.23919/INDIACom70271.2026.11525435.

[46] S. Mohamed, M.-T. Png, and W. Isaac, "Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence," Philosophy and Technology, vol. 33, pp. 659–684, 2020.

[47] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2021, pp. 610–623.

[48] S. R. Carroll et al., "The CARE principles for Indigenous data governance," Data Science Journal, vol. 19, no. 1, p. 43, 2020.

[49] J. Anderson and K. Christen, "Traditional Knowledge (TK) Labels," Local Contexts. [Online]. Available:

[50] Coalition for Content Provenance and Authenticity, "C2PA Technical Specification, Version 2.1," 2024. [Online]. Available:

[51] M. Mitchell et al., "Model cards for model reporting," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAT)*, 2019, pp. 220–229.

[52] T. Gebru et al., "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021.

[53] M. Doerr, "The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata," AI Magazine, vol. 24, no. 3, pp. 75–92, 2003.

[54] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[55] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," arXiv preprint arXiv:2212.08073, 2022.