DISEGNO

IX/01
journal of design culture

_Homogenised Heritage: AI and Central Europe

HOMOGENISED HERITAGE: AI AND CENTRAL EUROPE

THE IMPACT OF AI ON LOW-RESOURCE LANGUAGES AND VISUAL CULTURES IN THE VISEGRAD COUNTRIES

Disegno

Journal of Design Culture
Double-blind peer-reviewed, open access scholarly journal

Aims and Scope

Disegno publishes original research papers, essays, and reviews on all aspects of design cultures. We understand the notion of design culture as resolutely broad: our aim is to freely discuss the designed environment as mutually intertwined strands of sociocultural products, practices, and discourses. This attitude traverses the disciplinary boundaries between art, design, and visual culture and is therefore open to all themes related to sociocultural creativity and innovation. Our post-disciplinary endeavour welcomes intellectual contributions from all members of different design cultures. Besides providing a lively platform for debating issues of design culture, our specific aim is to consolidate and enhance the emerging field of design culture studies in the Central European academia by providing criticism of fundamental biases and misleading cultural imprinting with respect to the field of design.

All research papers published in Disegno undergo a rigorous double-blind peer review process.
This journal does not charge APCs or submission charges.

Contact: *Moholy-Nagy University of Art and Design*
H-1121 Budapest, Zugligeti út 9–25.
disegno@mome.hu

The full content of Disegno can be accessed online: ***disegno.mome.hu***

Published by: *Csaba Kovács*
Publisher: Moholy-Nagy University of Art and Design, 1121 Budapest, Zugligeti út 9-25.

ISSN: *2064-7778 (print)* ***ISSN:*** *2416-156X (online)*

MOHOLY-NAGY
művészeti egyetem
university of art and
design budapest

Contents

EPISTEMIC CULTURAL FLATTENING IN GENERATIVE VISUAL AI:

BENCHMARKING HUNGARIAN HERITAGE AND DESIGNING A V4 PATH TOWARD CULTURALLY AWARE TEXT-TO-VIDEO

Brigitta Iványi-Bitter, Tibor Bacsi, and Szilárd Szakács

DISEGNO_IX/01_HOMOGENISED HERITAGE: AI AND CENTRAL EUROPE

ABSTRACT

Generative image systems increasingly shape how culture becomes visible in design workflows and heritage interpretation. Their outputs often achieve technical plausibility while offering limited support for validating cultural provenance, shaping how synthetic images circulate as cultural references. This article introduces Epistemic Cultural Flattening (ECF) and an Epistemic Interpretive Framework (EIF) to distinguish structural performance from epistemic readability and to describe reductions of culture-specific legibility under globally dominant visual templates. The study operationalizes EIF through a cultural fidelity benchmark rating generated images by cultural fit, stylistic accuracy, and technical quality. It uses a Hungarian heritage benchmark set within a cross-cultural comparative corpus and compares outputs from four diffusion-based generators.

The article proposes an ECF failure-mode typology that makes cultural flattening visually legible. It also outlines a V4-oriented workflow for culturally aware text-to-video, integrating GLAM sourcing, multilingual metadata, controlled model adaptation, and expert review for low-resource cultures in Central Europe.

#epistemic cultural flattening; #cultural fidelity; #benchmark; #generative AI; #Hungarian visual heritage; #low-resource languages; #GLAM; #V4; text-to-video

https://doi.org/10.21096/disegno_2025_1bib

1. INTRODUCTION

1.1 The problem: high technical quality, limited cultural fidelity

Over the past few years, generative AI has shifted from a novelty to an everyday mediator of visual culture. Image generators now routinely support design prototyping, historical illustration, place simulation, and the on-demand visualisation of cultural references, reinforcing a broader condition in which cultural production and circulation increasingly operate through algorithmic infrastructures (Striphas 2015). The apparent fluency of these systems sustains a persistent paradox: an image can achieve high technical quality—sharp, coherent, visually persuasive—while cultural provenance and stylistic accuracy remain unstable. It can satisfy generic expectations of what folk embroidery, historic painting, or Central European architecture "should" look like, while producing weak support for recognition within the culture referenced by the prompt.

This gap matters because cultural meaning resides in the codes through which a subject becomes readable—ornamental grammar, material conventions, typographic habits, and historically situated stylistic rules (Hall 1997). When these cues shift toward globally dominant templates—generic "European old town" aesthetics, interchangeable "Eastern folk" ornament, or an accidental amalgam of period styles—validation of cultural provenance and stylistic accuracy becomes essential. The output then functions as a persuasive substitute whose authority arises from visual polish and coherence rather than from culturally situated evidence, a dynamic that becomes especially consequential as AI aesthetics normalise a distinctive regime of synthetic plausibility (Manovich 2019).

Seen from a design and cultural heritage perspective, this paradox follows from how generative systems are built and evaluated. Cultural legibility is mediated by infrastructures: training datasets and their uneven geographies of visibility; metadata and naming systems that shape what becomes learnable; prompt languages and translation layers that compress culturally specific terms; and evaluation regimes that reward coherence while leaving provenance weakly represented (Crawford 2021). In low-resource cultural contexts, including the Visegrad region, these infrastructures privilege what circulates widely through platformed extraction and large-scale data capture, so cultural requests get resolved through statistically dominant priors rather than through locally dense

reference structures (Couldry and Mejias 2019). Search engines and platforms further reinforce this dynamic by shaping discoverability and salience through ranking, categorisation, and moderation logics that structure what appears culturally retrievable at scale (Gillespie 2018). What emerges is a systematic tendency toward cultural flattening: outputs that look generally correct but whose culturally situated reading relies on cues that remain unevenly supported.

Ania Malinowska, in her contribution to this issue, conceptualises this structural dynamic as *AI assimilationism*, in which local aesthetics gain visibility through dominant norms and infrastructures shaped by English-language defaults and platformed hierarchies; her triad of linguistic standardisation, economies of visibility, and misidentification provides a close theoretical match to the infrastructural mechanisms traced here.

1.2 Research questions and contributions

This article approaches cultural fidelity in generative images as a design and cultural heritage problem with measurable outcomes. It frames the problem through three research questions that focus on measurement, empirical diagnosis, and design intervention.

RQ1: How does the benchmark measure divergence between technical quality and culturally situated recognisability in generative images?

RQ2: Which recurring failure modes characterise this divergence in Hungarian cultural heritage imagery across domains (fine art, folk art, and architecture)?

RQ3: How does this diagnosis shape a design workflow for generative tools that support culturally specific outputs across the V4 countries, including text-to-video applications?

1.3 Contributions

This article offers five contributions that connect conceptual framing, measurement, empirical evidence, and design intervention:

Conceptual: *epistemic cultural flattening* (ECF) is introduced as a term for culturally plausible-looking outputs that lose provenance-specific legibility and this phenomenon is framed through an *epistemic interpretive framework* (EIF) that separates structural performance from epistemic readability.

Methodological: A cultural fidelity benchmark is proposed that evaluates AI-generated images along three complementary dimensions (cultural fit, stylistic accuracy, and technical quality) to support systematic comparison across models, domains, and cultures.

Empirical: Results from a Hungarian heritage benchmark set situated within a cross-culture comparative corpus are reported using multiple image generators to map how cultural fidelity varies by domain (fine art, folk art, and architecture) and by model family.

Analytical: An ECF failure-mode typology is developed that makes cultural flattening visually legible through recurring patterns such as generic substitution, motif drift, semantic collapse of local terms, and style–reference decoupling effects.

Design: The diagnosis is translated into a V4-oriented workflow for culturally aware generative tools, extending image generation to text-to-video and grounded in GLAM collaboration, multilingual metadata practices, model fine-tuning, and iterative expert evaluation.

1.4 Paper roadmap

This paper is structured as follows. Section 2 introduces ECF and the EIF, and it differentiates these concepts from adjacent terms commonly used to describe generative error and cultural bias. Section 3 presents the cultural fidelity benchmark and the empirical material that supports it, including the benchmark dimensions, the dataset design, the prompting strategy, the reference-image grounding approach, and the evaluation procedure. Section 4 reports the empirical results, with a focus on how cultural fit and stylistic accuracy vary in relation to technical quality across domains and model families within the Hungarian heritage benchmark set and the cross-culture comparative corpus.

Building on these findings, section 5 develops an ECF failure-mode typology that describes how cultural flattening becomes visible at the level of motifs, styles, and provenance cues. Section 6 translates the diagnostic insights into a V4-oriented design workflow for culturally aware generative tools, extending the logic of the benchmark toward text-to-video development through GLAM collaboration, multilingual metadata practices, controlled model adaptation, and iterative expert evaluation. Section 7 discusses implications in three domains: design research and evaluation practice, GLAM institutions and cultural policy, and V4 toolmaking for creative and educational use. Section 8 concludes the paper by summarising the main contributions and outlining limitations alongside future research directions.

The next section establishes the conceptual vocabulary that frames the benchmark, guides the empirical analysis, and supports the design pathway developed in the second half of the paper.

2. CONCEPTUAL FRAMING: DEFINING AND DIFFERENTIATING ECF

2.1 Epistemic Interpretive Framework (EIF): two layers of evaluation

This article uses an EIF to separate two forms of performance that often collapse into a single idea of "image quality" in everyday use. The EIF treats generative outputs as cultural representations that require evaluation on two analytically distinct layers: structural performance and epistemic readability.

Layer A: *Structural performance* captures the technical and compositional competence of a generated image, independent of its cultural attribution. It captures qualities that remain largely transferable across contexts, such as technical clarity, compositional coherence, artifact handling, and overall visual plausibility. In practical terms, this layer

corresponds to the kind of acceptable quality that becomes visible through resolution, lighting consistency, surface detail.

Layer B: *Epistemic readability* describes an image's capacity to carry culturally situated meaning in a way that supports recognition by informed observers. This layer centres on *cultural fit* and *stylistic accuracy*. Cultural fit concerns whether the output activates culture-specific signifiers that make the depicted object, place, or tradition legible within the referenced cultural context. Stylistic accuracy concerns whether the output follows the relevant aesthetic and material conventions, such as ornamental grammar, formal structures, craft logics, architectural typologies, or art-historical registers associated with the referenced domain. Epistemic readability therefore measures the extent to which an output sustains validation of cultural fit and stylistic accuracy within the referenced context.

Within EIF, ECF appears as a patterned divergence between these two layers: An output can score highly on structural performance while showing low epistemic readability. This divergence matters because it changes how images function in cultural circulation. High structural performance increases persuasive force, while reduced epistemic readability shifts cultural specificity toward generic templates. EIF therefore makes it possible to diagnose cultural failure modes that remain hidden when evaluation focuses on technical quality alone.

The ECF phenomenon is close to what Markova (in her article in this issue) frames as a parallel risk through *collective vulnerability*, where profit-driven AI development flattens cultural diversity toward a computational mean, a framing that complements ECF by linking output-level divergence to broader power structures.

2.2 Definition: Epistemic cultural flattening (ECF)

ECF describes a patterned shift in generative outputs in which culturally specific meaning becomes less distinguishable, even as visual coherence remains high. Within the EIF, ECF names the divergence between structural performance and epistemic readability: an image can score highly on technical and compositional quality while offering limited support for validation as a culturally situated representation. ECF therefore captures a form of representational homogenisation in which outputs default toward globally dominant templates and broadly legible cues, and in doing so reduce the visibility of culture-specific ornamental grammar, material conventions, and historically situated styles.

Diagnostic signals appear as score divergence between technical quality and cultural fit and stylistic accuracy. Visual form often stabilises through template substitution and provenance thinning.

In this issue, Janda describes a closely aligned phenomenon to ECF as *regional invisibility*, where culturally dense low-resource places become reconstructed through globally legible templates, producing subtle drift through normalisation, typological substitution, and shifts toward universally plausible aesthetics.

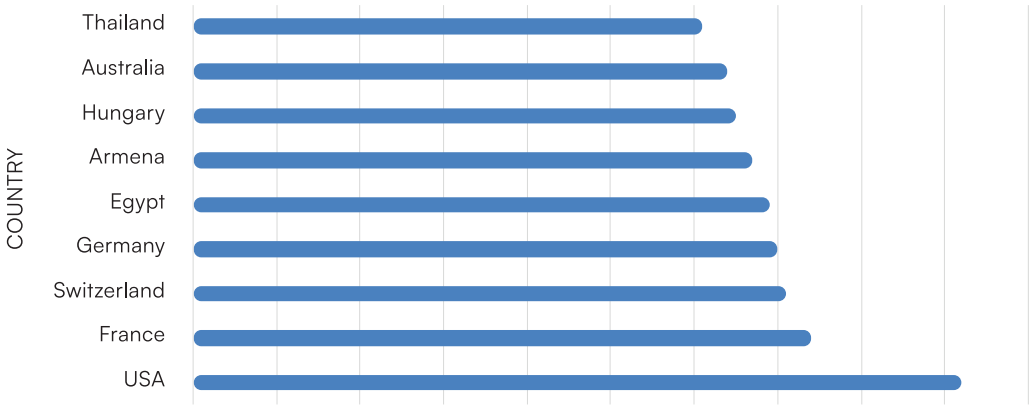


FIGURE 1. *Situating ECF among adjacent concepts*

ECF gains analytical precision through its contrast with terms that circulate widely in public and technical discussions of generative AI. These labels often describe genuine phenomena, while ECF focuses on a specific representational pattern: the reduction of culture-specific legibility under globally dominant visual templates. Korinek in his contribution to this issue, frames this asymmetry as a Central European condition: small-language environments enter into collaboration with global generative systems from an unequal position and the process amplifies tension between local experience and algorithmically preferred forms of language and imagery. This perspective strengthens the conceptual scope of ECF as an output-level pattern that remains structurally linked to cultural position within global data environments.

ECF and hallucination

In technical discussion of generative AI "hallucination" typically denotes outputs that contain implausible elements. ECF has a different regime of error: outputs often remain visually coherent and aesthetically plausible, while cultural attribution relies on substitution. The image reads as something like the requested cultural tradition, place, or artwork, and it achieves plausibility through generic cues rather than through culture-specific references. In this issue Krzykawski's critique of *AI parlance* complements this differentiation work by foregrounding how everyday terms such as hallucination shape public understanding and flatten conceptual distinctions that support critical evaluation.

ECF and bias

"Bias" commonly names systematic disparities that affect social groups, including stereotyping and unequal treatment across gender, race, or ethnicity. ECF complements this conversation by foregrounding a representational dimension that concerns cultural legibility, especially in heritage contexts. It highlights how ornamental grammar, typological specificity, and historically situated stylistic registers become diluted when models resolve prompts through dominant templates.

ECF and localisation

"Localisation" typically emphasises geographic placement and language adaptation, often evaluated through the extent to which a model depicts the correct place or renders text in the appropriate language. ECF addresses a broader representational mechanism: a scene can appear "Eastern European" while still losing the specific cues that support Hungarian, Polish, Czech, or Slovak cultural attribution. Cultural specificity depends on more than place names; it depends on typologies, materials, style conventions, ornamental grammar, and craft logic.

ECF and style transfer

"Style transfer" normally focuses on mapping one visual style onto another image or subject, and evaluation often concentrates on aesthetic resemblance. ECF speaks to the relationship between style and provenance. An output can match a plausible stylistic register while cultural attribution remains weak when provenance anchors (like regional typology) remain underrepresented in the training data.

ECF and dataset scarcity

Dataset scarcity describes uneven representation within training corpora, a condition that shapes what models can learn. ECF names the epistemic outcome that follows from such conditions: cultural requests get resolved through the widely available templates, and cultural distinctiveness shifts toward generalised visual forms. Dataset scarcity functions as a driver; ECF describes the representational effect visible in outputs and measurable through evaluation.

3. METHODS: THE CULTURAL FIDELITY BENCHMARK

3.1 Benchmark logic

This study operationalises cultural fidelity through a benchmark instrument that separates three evaluative dimensions: *cultural fit*, *stylistic accuracy*, and *technical quality*. The instrument responds to a recurring challenge in the assessment of generative images: outputs often achieve strong technical quality, and this competence increases their persuasive force in cultural circulation. Evaluation therefore benefits from a structure that distinguishes visual coherence from culturally situated legibility.

Cultural fit captures whether an output belongs to the referenced cultural context in a way that supports recognition by informed observers. It concerns the presence of culture-specific cues, such as typologies, iconographic conventions, and ornamental grammar, that enable viewers to attribute the image to the intended culture rather than to a generic proxy. Cultural fit therefore addresses the question of cultural attribution: the relationship between the prompt's cultural reference and the image's culturally situated readability.

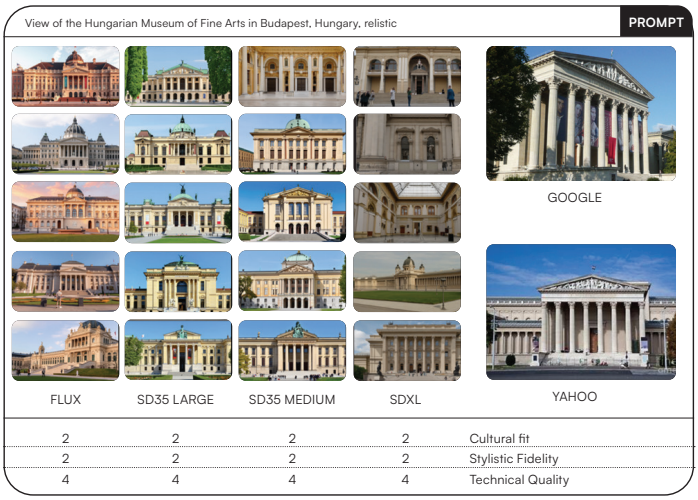


FIGURE 2. *Cultural fidelity benchmark instrument (three domain exemplars). The benchmark evaluates AI-generated images along three dimensions (cultural fit, stylistic accuracy, and technical quality) using a consistent 1–7 Likert scale. The three cards illustrate how the same evaluation backbone applies across domains while the interpretive focus shifts with the three categories.*

Stylistic accuracy captures whether the output follows the relevant aesthetic and material conventions associated with the referenced domain. It concerns formal and material coherence within a tradition: the exact structure of ornament, characteristic colour relations, compositional logic, craft constraints, and domain-specific visual registers. Stylistic accuracy therefore focuses on how faithfully the output aligns with the stylistic rules that shape a tradition's internal visual logic.

Technical quality captures structural performance at the level of the image as a rendered artifact. It concerns clarity, resolution, compositional stability, texture handling, and the overall absence of distracting visual elements that compromise visual coherence. Technical quality supports cross-model comparison because it remains interpretable across domains and cultures, while cultural fit and stylistic accuracy remain context-dependent.

Together, the three dimensions support an interpretable diagnosis of ECF. High technical quality can coexist with low cultural fit or stylistic accuracy, and this divergence provides a measurable trace of ECF. The benchmark therefore functions as a practical instrument for comparative evaluation across models, domains, and cultures, while also supporting qualitative interpretation through exemplar images. Benchmarking traditions in algorithmic accountability show how structured evaluation reveals systematic performance gaps and guides intervention, a logic that supports the cultural fidelity benchmark developed here (Buolamwini and Gebru 2018).

3.2 Dataset construction with Hungarian heritage as anchor

The benchmark builds on a comparative image-generation corpus designed to support controlled cross-cultural analysis while keeping Hungarian visual heritage at the centre of interpretation. The dataset comprises approximately 900 generated images produced with four diffusion image generators (Stable Diffusion XL, Stable Diffusion 3.5 Large, Stable Diffusion 3.5 Medium, and Flux Schnell) across ten cultures and three domains: architecture, fine art, and folk art. While the comparative corpus targets ten cultures, the results reported here draw on the eight for which evaluation data was complete at the time of analysis.

The benchmark uses countries as proxies for cultural contexts, a simplification that treats each national dataset as representative of a broadly identifiable visual heritage tradition. This operationalisation supports controlled comparison while recognising that cultural production within any country is internally diverse and that national boundaries do not map neatly onto cultural boundaries.

The selection of cultures functions as a set of comparative positions in global data hierarchies, spanning contexts with high digital resources (the United States, France, Germany, Switzerland, Australia) and contexts with lower digital resources (Hungary, Bangladesh, Thailand). One additional low-resource culture (Egypt) was targeted but not yet evaluated at the time of analysis.

Since this research began with an inquiry into misrepresentation in Hungarian heritage, the Hungarian dataset serves as the anchor case. The benchmark targets heritage objects and references that rely on culture-specific cues for recognition, including ornamental grammar, material conventions, regionally specific typologies, and art-historical provenance. The two comparator groups support interpretation in two ways: high-resource contexts function as baselines for strong learnability within global training corpora, while lower-resource contexts clarify how cultural specificity behaves under sparse representation conditions.

The corpus is organised through a domain-based prompt set that aligns the three heritage categories with comparable prompt structures across cultures. Each culture receives prompts in each domain, and each prompt is executed across all four generators to enable model-by-model comparison under identical textual conditions.

This design treats cultural representation as a pattern that emerges across repeated prompts rather than as an isolated anecdote: multiple outputs per prompt and per model support the identification of recurring representational logics tied to a given culture and domain.

To support cross-culture comparability, the prompt structures remain standardised while allowing culture-specific references to enter through culturally situated objects.

3.3 Generation systems, grounding, evaluation design, and analysis approach

Cross-model comparison matters because cultural fidelity rarely behaves as a simple function of technical advancement: models differ in rendering competence, stylistic priors, and text–image alignment, and these differences shape how culturally specific prompts resolve into visual form. A shared prompt set executed across multiple generators therefore supports two complementary readings: model-level differences in cultural fidelity and domain-level patterns that remain stable across models, both of which inform the diagnosis of epistemic cultural flattening.

To support culturally situated evaluation, each prompt was paired with two reference images sourced through search engines. These references function as grounding anchors rather than as definitive truth claims. Their role is practical and comparative: they provide evaluators with a memory aid for the culturally referenced object, place, or work, and they render visible the visibility regimes that structure online access to heritage imagery. Reference selection therefore carries diagnostic value because search results reflect platformed hierarchies, dominant iconographies, and metadata conventions that shape what becomes culturally retrievable at scale, including ranking and categorisation effects that organise visibility in platform environments (Gillespie 2018). In this setting, reference images support situated judgement by assisting evaluators in assessing cultural fit and stylistic accuracy while keeping the evaluation open to plurality within a tradition.

This methodological choice also aligns with Markova's observation, stated in her article in this issue, that research on Central European cultural alignment remains limited and that artistic and design research practices offer productive approaches for developing insight. Benchmark-based evaluation therefore operates as an interpretive instrument that combines structured scoring with culturally situated reference points to make patterns of alignment and flattening empirically legible.

Evaluation relied on culturally informed judgement. Evaluators were therefore recruited for their familiarity with relevant cultural contexts and heritage domains, and expertise was operationalised as the ability to recognise culture-specific cues and articulate domain conventions in folk art, fine art, and architecture. The study design used balanced presentation logic to support comparability and reduce fatigue effects. Prompt–model combinations were distributed so that evaluators encountered a controlled mix of domains and systems, and the assignment followed a structured balancing scheme (including Latin-square style distribution) to stabilise order effects across the full set. In addition to numeric ratings of cultural fit, stylistic accuracy, and technical quality, the protocol included brief qualitative notes. These notes function as interpretive traces that connect aggregate scores to recurrent visual patterns, especially in cases where evaluators identify substitution, genericisation, or weakened provenance cues.

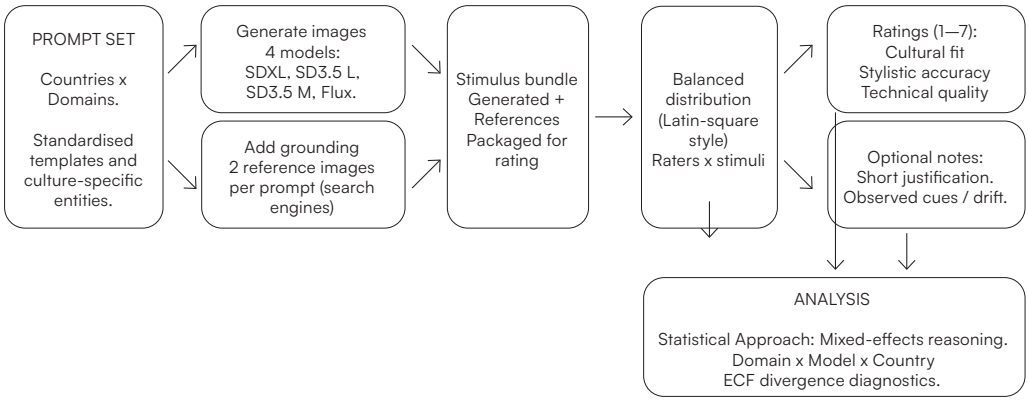
Inter-rater agreement was assessed using Krippendorff's alpha with ordinal distance. Across the three benchmark dimensions, agreement fell in the range $\alpha = 0.20$–0.29 (cultural fit: $\alpha = 0.29$; stylistic accuracy: $\alpha = 0.20$; technical quality: $\alpha = 0.27$). These values fall below the conventional threshold of $\alpha \geq 0.67$ recommended for definitive conclusions (Krippendorff 2004). However, moderate-to-low agreement is characteristic of tasks that involve culturally situated aesthetic judgment rather than factual classification. The benchmark does not claim ground-truth coding: it treats ratings as expert perceptions of culturally referenced outputs, where legitimate variation across evaluators reflects the interpretive plurality inherent in heritage assessment. With approximately three raters per culture, the design supports directional and benchmark-level claims rather than fine-grained pairwise inference.

Analysis treats the dataset as a multi-level structure in which prompts, models, evaluators, cultures, and domains contribute systematic variation. This structure supports mixed-effects reasoning: prompts differ in difficulty, evaluators differ in calibration, and models differ in priors, so comparison benefits from an approach that accounts for nested sources of variance. The analysis therefore centres on two perspectives: a Hungary-centred reading that treats Hungarian heritage as the anchor case, and a-cross-culture comparative reading that situates Hungary within broader visibility gradients. Results are reported in a form that remains readable for design research audiences through domain-level contrasts, model-level contrasts, and the divergence pattern that operationalises ECF as a gap between technical quality and epistemic readability.

Ratings were analysed using linear mixed-effects models (LMMs) with model, domain, and cultural resource level (high vs. low) as fixed effects and participant and prompt as crossed random intercepts. This specification accounts for systematic differences in rater calibration and prompt difficulty while estimating the effects of interest. Responses marked "cannot evaluate" (126 of 3,218 total responses, approximately 4%) were excluded from statistical analysis; these predominantly occurred in prompts requiring familiarity with specific heritage objects outside the evaluator's domain of expertise. An order-effects check confirmed no substantive position effects across dimensions (all $p > .05$), supporting the effectiveness of the randomised presentation design.

Methodologically, Janda's Total Distortion Score approach (published in this issue) complements this benchmark logic by treating drift as structured and repeatable and by coding variables of regional distortion across systems, strengthening comparative reading of low-resource visual contexts.

FIGURE 3. *Study pipeline schematic*



4. RESULTS: DIAGNOSING ECF IN THE HUNGARIAN HERITAGE SUBSET

4.1 Core pattern: technical success can coexist with reduced cultural fidelity

Across the Hungarian heritage benchmark set, results show a stable divergence between structural performance and epistemic readability. Generated images often achieve high scores on *technical quality* (clear rendering, coherent composition, and plausible surfaces) while evaluators assign lower scores to *cultural fit* and *stylistic accuracy*.

The magnitude of this divergence is confirmed by effect-size analysis. Across the comparative corpus, high-resource cultures received significantly higher ratings than low-resource cultures on all three benchmark dimensions: cultural fit ($d = 0.43$, $p < .001$), stylistic accuracy ($d = 0.64$, $p < .001$), and technical quality ($d = 0.47$, $p < .001$), corresponding to medium effect sizes by conventional benchmarks (Cohen 1988). The strongest divergence appears in stylistic accuracy, where domain-specific visual conventions

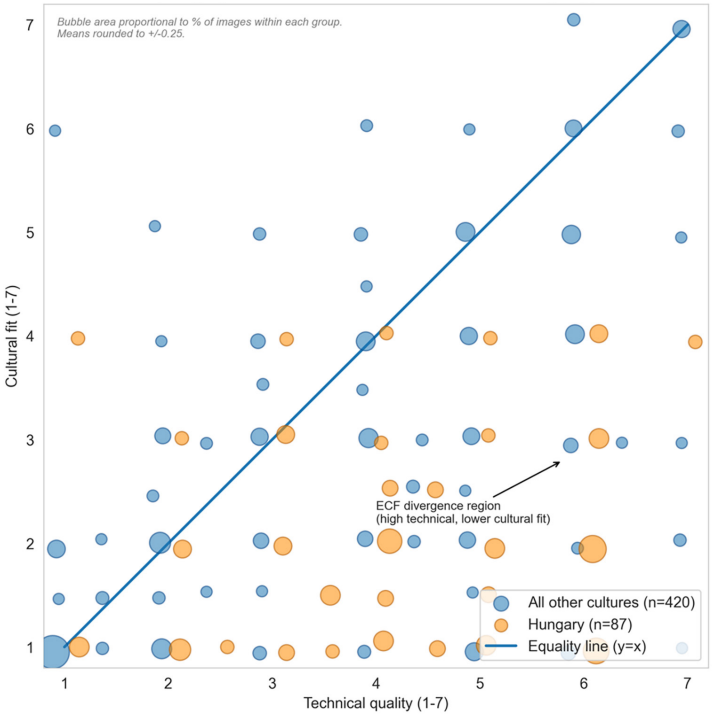


FIGURE 4. *Divergence plot: technical quality vs. cultural fit (Hungary vs. the other countries; bubble area represents percentage of group).*

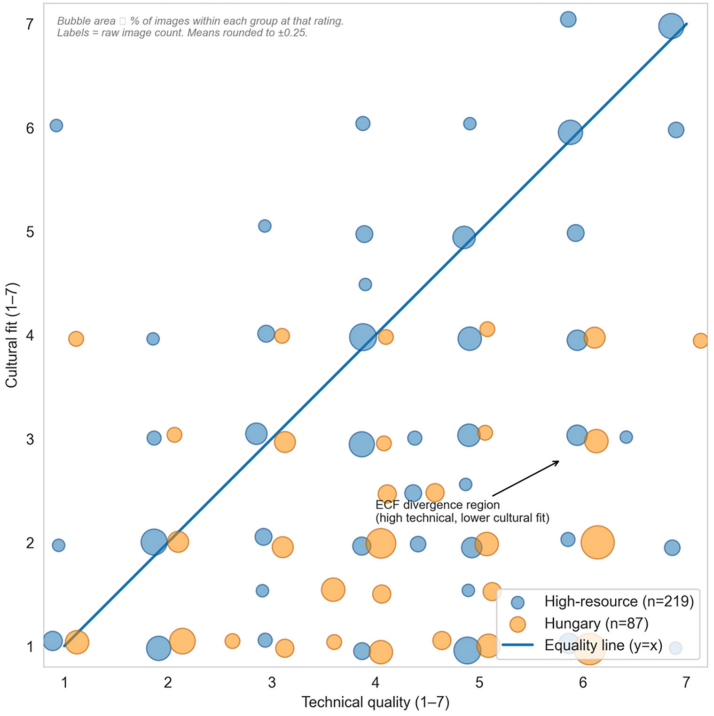


FIGURE 5. *Divergence plot: technical quality vs. cultural fit (Hungary vs. high-resource; bubble area represents percentage of group).*

are most dependent on culture-specific training coverage. That medium effect sizes emerge consistently across all three dimensions, rather than concentrating in cultural fit alone, strengthens the interpretation of ECF as a systematic representational pattern rather than an isolated scoring artifact.

The divergence becomes especially salient in prompts that depend on culturally specific anchors rather than globally common objects. In these cases, the generated image frequently presents a visually acceptable proxy: a genericised "folk" surface for folk art, a broadly "European" architectural scene for place-based heritage, or a historically plausible painting register for fine art. Technical quality supports the plausibility of these proxies, while cultural fit and stylistic accuracy depend on finer-grained cues (ornamental grammar, material conventions, typological specificity, and provenance anchors) that shape culturally situated validation. ECF therefore appears less as a breakdown of image generation and more as a patterned shift in representational strategy: the output prioritises globally legible templates and stabilises cultural meaning through substitution rather than through culture-specific evidence.

Within the Hungarian subset (n = 87 images), the divergence pattern is pronounced: mean technical quality (M = 4.22, SD = 1.55) exceeds mean cultural fit (M = 1.92, SD = 0.94) by 2.30 points.

4.2 Domain differences

Across domains, the results form a clear visibility pattern shaped by the interaction of cultural resource level and heritage type (figure X). In high-resource contexts, architecture tends to remain comparatively stable because place depiction can be assembled from widely circulating photographic templates and globally legible built-environment cues; cultural attribution often holds at the level of recognisability. In the same high-resource contexts, folk and fine art achieve a partial fidelity: models often reproduce a plausible aesthetic register while fine-grained provenance cues and domain-specific ornamental grammar remains uneven, producing outputs that support broad recognition rather than through tradition-specific evidence. Folk prompts also function as representational triggers: Keszeg's contribution to this issue demonstrates how *ethnicising bias* structures the visual grammar through which generative AI systems render folk culture.

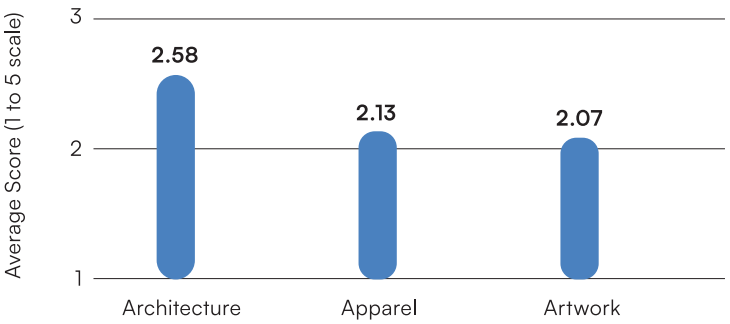


FIGURE 6. *Average AI cultural fidelity score by category.*

In low-resource contexts, the pattern shifts in two distinct ways. Architecture remains visible yet tends toward distortion: outputs stabilise into generic "European" scenery or interchangeable urban typologies, and place-specific anchors weaken, producing cultural attribution that requires additional validation. Folk and fine art show the strongest compression of cultural specificity and therefore approach invisibility at the level that matters for heritage reading: models sustain surface plausibility while the cues that enable culturally situated recognition, like regionally specific typologies, craft constraints, and ornamental grammar, thin out. This comparison summarises how ECF intensifies when cultural requests demand high-resolution specificity and when training data offers limited coverage of the relevant provenance anchors. The cross-domain visibility pattern resonates with Malinowska's contribution to this issue, particularly her account of *economies of visibility*, in which forms that already circulate widely gain further amplification through platformed selection.

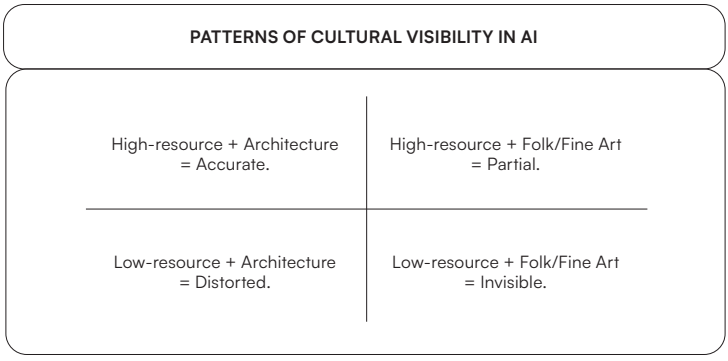


FIGURE 7. *Patterns of Cultural Visibility in AI.*

4.3 Model differences

The Hungary subset indicates small but consistent differences in mean cultural-fit ratings across the four image-generation systems. SDXL receives the highest mean cultural-fit score (M = 2.78), followed by Stable Diffusion 3.5 Medium (M = 2.64) and Stable Diffusion 3.5 Large (M = 2.59), while Flux Schnell receives the lowest mean rating (M = 2.49). Taken together, these values suggest a modest descriptive ordering among the models rather than a sharply differentiated hierarchy.

This ordering remains useful for comparative interpretation. It helps situate the kinds of images that different systems tend to stabilise under identical prompts and shows that some models produce outputs that are judged as somewhat more culturally fitting than others. At the same time, the relatively narrow spread between the mean values calls for caution. The differences shown in figure 8 should therefore be read as descriptive contrasts in average ratings, not as evidence of large separation in model performance.

Most importantly, the model comparison leaves the core ECF pattern intact. Even where one system performs somewhat better than another on average cultural fit, the broader structure of ECF remains visible across

all four systems. Improvements at the model level may raise overall capability, but they do not resolve the more fundamental divergence between technical image generation and culturally grounded readability. In other words, model choice matters for relative performance, yet the persistence of ECF points to a wider infrastructural condition that shapes cultural visibility across systems rather than to a problem confined to any single model.

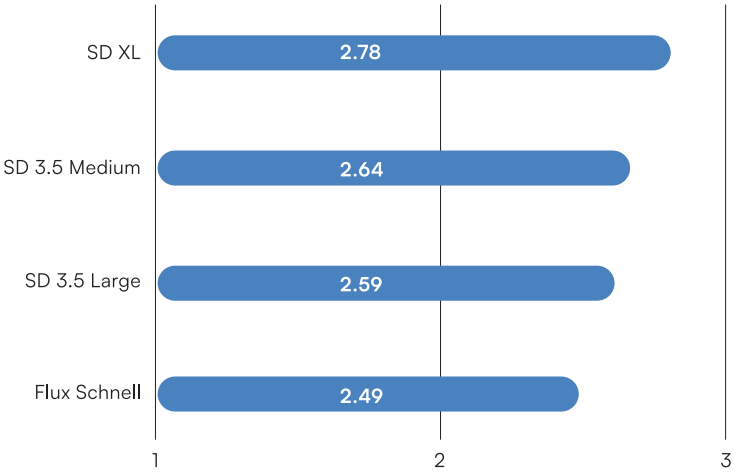


FIGURE 8. *Descriptive mean cultural-fit ratings across models for the Hungary subset.*

5. ECF FAILURE-MODE TYPOLOGY

Benchmark scores provide a necessary diagnosis of cultural fidelity. The benchmark identifies divergence between technical quality and epistemic readability, and it supports comparison across models, domains, and cultures. A typology adds a second layer of interpretive work: it translates measured outcomes into a design vocabulary that describes recurring representational problems.

First, this vocabulary makes ECF readable as a set of repeatable visual patterns, like substitutions, drifts, and reductions of provenance anchors, rather than as isolated errors. Second, it supports action. Designers, data curators, and tool developers benefit from terms that describe where cultural meaning collapses and how it collapses, because these descriptions guide both prompt design and dataset intervention. Typologies therefore function as bridge instruments: they connect quantitative evaluation with qualitative diagnosis, and they support iterative improvement through targeted tests, controlled comparisons, and domain-specific refinement.

The following typology describes six recurring ways in which ECF appears in generated images. Each mode names a representational mode, summarises the visual symptoms that make it recognisable, identifies the prompt conditions that tend to trigger it, and links the pattern to a benchmark across cultural fit, stylistic accuracy, and technical quality.

Generic-European substitution describes a resolution strategy in which a culture-specific request is satisfied through broadly legible "European" templates that support a general regional reading while narrowing culture-specific attribution. Visually, outputs converge on postcard-like urban textures, familiar rooflines, and standardised streetscapes, and they rely on scenic composition rather than on typological anchors and local material details. This mode appears frequently in architecture prompts framed as "view of…" and in folk prompts that specify "traditional" heritage without stronger material constraints. In the benchmark, the pattern typically presents as lowered cultural fit with a mild-to-moderate decrease in stylistic accuracy, alongside high technical quality. Keszeg's contribution to this issue provides a related cross-country reading of substitution as an *imaginary shift*, where outputs stabilise through Mitteleuropean, generalised Slavic, or Orientalised "Eastern" frameworks in response to prompt cues.

Ornamental drift describes outputs that present decorative patterning that reads as heritage ornament while the culture-specific ornamental grammar and craft logic remain unstable. Visually, motifs appear plausible yet reorganise into globally common floral geometry; colour relations move toward standardised palettes associated with generic folk aesthetics; and stitch logic or material behaviour reads as surface decoration rather than craft constraint. This mode concentrates in folk art prompts involving textiles, dress, embroidery, and decorative crafts. In the benchmark, stylistic accuracy tends to drop most clearly, cultural fit often follows with a smaller decrease, and technical quality remains high.

Semantic collapse of local terms describes a prompt–output shift in which culture-specific terms compress into a broader category label, steering the image toward generic object types and generalised heritage cues. Visually, named objects become category-level proxies, for example a traditional guba coat, Miska jug, or regional costume, and key identifiers lose strength while generic decorative cues increase. This mode appears often when prompts include Hungarian terms, diacritics, or regionally specific names, and when prompts combine Hungarian and English descriptors. In Keszeg's contribution to this issue, this mechanism aligns with *representational displacement under noisy labelling*, whereby locally specific visual forms are displaced by more widely available regional or generic repertoires. The benchmark typically records a decrease in cultural fit accompanied by a smaller decline in stylistic accuracy, with technical quality remaining high.

Anachronistic hybridisation describes outputs that integrate stylistic cues from multiple time periods into a single image, producing coherent scenes with unstable historical placement. Visual symptoms include garments that combine silhouettes and accessories from different eras, architectural depictions that mix façade motifs and material treatments associated with distinct periods, and fine art scenes that drift across

historical registers while maintaining an era-like look. This mode often arises in prompts that include dates, in historic architecture prompts, and in folk costume prompts framed as "realistic" without specifying a documentary register. In the benchmark, stylistic accuracy typically declines first, cultural fit follows with a smaller decrease, and technical quality remains high.

Style-source decoupling describes outputs that match a plausible period or genre register while source-specific anchors that support attribution to a named artwork, artist, or tradition remain weak. Visually, prompts naming artworks or artists yield images that "fit the era" while composition, iconography, and work-level identity drift; portrait, devotional, or plein-air scenes appear as genre-typical substitutes; and visual polish increases credibility while provenance cues thin out. This mode concentrates in fine art prompts that reference named works or artists from Hungarian art history. Benchmark scores typically show lowered cultural fit with a mild-to-moderate decline in stylistic accuracy, alongside high technical quality.

Locational blur in architecture describes outputs that produce plausible built-environment depictions while place-specific anchors that support landmark recognition and site attribution remain partial. Visually, landmarks resolve into generic historic façades or scenic city views; spatial context aligns with common tourist-photography conventions; and materials, ornamentation logic, and massing support a general regional reading. This mode appears frequently when prompts rely on a single place name or building name as the main constraint. The benchmark signature typically presents as lowered cultural fit and a smaller decline in stylistic accuracy, with technical quality remaining high.

Together, these six modes convert benchmark divergence into a practical design vocabulary.

FIGURE 9. *ECF failure-mode typology matrix.*

Failure mode	Core visual symptoms	Typical triggers	Benchmark signature
1. Generic-European Substitution	Generic European templates; interchangeable cues	Architecture 'view of…'; folk prompts with weak constraints	Cultural fit ↓ Style ↘ Tech ↗/high
2. Ornamental Drift	Decorative patterning; unstable ornamental grammar	Folk textiles, dress, decorative crafts	Style ↓ Cultural fit ↘ Tech ↗/high
3. Semantic Collapse	Named term → category proxy; object identity ambiguity	Local terms, diacritics, mixed HU/EN phrasing	Cultural fit ↓ Style ↘ Tech ↗/high
4. Anachronistic Hybridization	Mixed period cues; historical register drift	Dated prompts; historic costume/architecture	Style ↓ Cultural fit ↘ Tech ↗/high
5. Style–Source Decoupling	Era/genre mood replaces work-level anchors	Named artists/works; Hungarian art history prompts	Cultural fit ↓ Style ↘ Tech ↗/high
6. Locational Blur	Plausible scene; weak landmark/site anchors	Named buildings; place name as main constraint	Cultural fit ↓ Style ↘ Tech ↗/high

6. FROM DIAGNOSIS TO DESIGN: TOWARD A V4 CULTURALLY AWARE TEXT-TO-VIDEO WORKFLOW

6.1 Why video raises the stakes

The diagnosis of ECF gains additional urgency in the transition from image generation to video generation. Text-to-video amplifies cultural representation through temporal continuity, narrative structure, and embodied cues.

Temporal consistency raises the stakes. Video systems must maintain cultural cues across frames, and this requirement turns minor drift into a visible structural problem. Ornament, materials, typological anchors, and stylistic registers must persist as stable features rather than as accidental successes in single frames. Temporal coherence therefore functions as a stress test for cultural fidelity: a model that occasionally produces a culturally plausible still image can still yield a culturally unstable video when key cues fluctuate across time.

Narrative structure raises the stakes further. Text-to-video workflows typically embed visual synthesis within short scripts or prompts that imply roles, settings, and causal sequences. These scripts activate templates for "what usually happens," and such templates often carry stereotyping pressure. Cultural specificity in heritage contexts relies on situated relations between objects, places, gestures, and social practices, while generative narrative defaults often rely on globally dominant story grammars. As a result, video generation increases the risk that cultural meaning becomes organised through familiar narrative clichés that displace local history, regional nuance, and context-specific social imagination.

Embodied cues raise the stakes a third time. Cultural recognition often relies on how bodies move through space, how garments sit and behave on bodies, how tools are handled, how rituals unfold, and how built environments structure everyday action. These embodied cues matter in V4 heritage contexts because they carry tacit knowledge that remains difficult to encode as isolated visual tokens. Video therefore shifts evaluation toward performative fidelity: the relationship between dress and movement, between craft and gesture, and between architecture and everyday use. This shift expands the benchmark logic beyond static representation and positions culturally aware text-to-video as a design challenge in which temporal stability, narrative choice, and cultural knowledge function as core variables.

Kořínek's (forthcoming) discussion of video work in which *temporal trace* appears as a visible imprint of a specific stage of the technology's development reinforces the value of temporal cultural fidelity as a future evaluation dimension.

6.2 Proposed V4 workflow

The diagnosis of ECF supports a practical intervention path: a staged V4 workflow that integrates cultural governance with model development. The workflow treats cultural fidelity as an engineered property shaped by institutional partnerships, multilingual description practices, controlled model adaptation, and iterative evaluation.

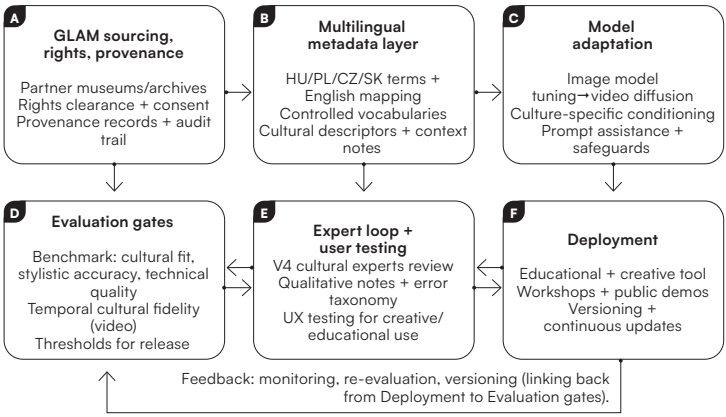


FIGURE 10. *Proposed V4 workflow (design + governance stack) toward culturally aware text-to-video.*

Stage A focuses on GLAM sourcing, rights, and provenance. The process begins through partnerships with museums, archives, and heritage organisations that hold regionally specific collections in Hungary, Poland, Czechia, and Slovakia. This stage structures permissions, consent, and documentation practices, and it establishes a provenance record for each asset, forming an audit trail that supports accountability.

Stage B builds a multilingual metadata layer that functions as a cultural interface. Heritage data gains usability for generative systems through controlled vocabularies and descriptive fields in HU/PL/CZ/SK with an English mapping that supports cross-country comparison and prompt tooling. This layer encodes typologies, ornamental grammar descriptors, material conventions, period registers, and contextual notes that assist both training and evaluation, and it strengthens cultural specificity at the level of language.

Stage C translates the dataset into model adaptation, progressing from image to video. The workflow uses image generation as the first stabilisation step, because images provide fast iteration cycles for cultural fidelity. Cultural conditioning then extends toward video diffusion once image-level performance supports validation.

Stage D introduces evaluation gates that combine the benchmark with a temporal extension for video. The benchmark dimensions (cultural fit, stylistic accuracy, and technical quality) serve as a release gate for image outputs, and the same logic extends toward video through temporal cultural fidelity criteria. Temporal cultural fidelity evaluates stability of key cues across frames, continuity of ornament and material behaviour, and coherence of typological anchors in motion.

Stage E operationalises an iterative expert loop and user testing. V4 cultural archive experts review outputs using both scores and brief qualitative notes aligned with the ECF typology, enabling targeted corrections in prompting, metadata, and adaptation strategy. User testing then evaluates whether the tool supports creative and educational use, including how users interpret cultural cues and how interface choices shape cultural outcomes.

Stage F deploys the system as an educational and creative tool with continuous versioning. Deployment includes workshops, public demonstrations alongside monitoring and periodic re-evaluation by data curators that supports iterative updates and governance over time.

6.3 How the workflow specifically targets ECF

The proposed V4 workflow targets ECF through a shift in what the system treats as learnable cultural evidence. ECF arises when models resolve culturally specific requests through globally dominant templates because these templates offer high statistical stability and broad visual legibility. The workflow intervenes by increasing *epistemic density*, meaning the availability, precision, and internal consistency of culture-specific cues that support validation of cultural provenance and stylistic accuracy.

Epistemic density can grow through governance choices. Rights-cleared GLAM sourcing and provenance documentation establish traceable cultural reference points, and this traceability supports accountability in model development and public deployment. Epistemic density then grows through the multilingual metadata layer, which functions as a cultural interface that translates heritage knowledge into structured descriptors. Controlled vocabularies, bilingual mappings, and context notes supply the model with richer anchors than generic labels.

The workflow also targets ECF by making evaluation an active design component. Evaluation gates translate cultural fidelity into release criteria during development. Expert review and qualitative notes then connect failure modes to actionable causes (prompt structure, metadata gaps, and conditioning weaknesses) supporting targeted iteration. This approach reduces misrecognition through design choices that stabilise provenance anchors and support culturally situated validation. The workflow also responds to Kuchta's observation in this issue that institutional virtual archives can feed AI image generators, so archival omissions and tagging structures can shape downstream cultural legibility in generated outputs.

7. IMPLICATIONS

7.1 Implications for design research and evaluation practice

The findings position cultural fidelity as a core concern for design research that engages generative systems as cultural infrastructures. Evaluation practices that emphasise visual coherence and technical polish capture only one layer of performance, while cultural meaning remains mediated by provenance anchors, ornamental grammar, typological specificity, and historically situated stylistic registers. Treating cultural fidelity as a first-class metric therefore expands evaluation beyond "does it look good" toward "does it support culturally situated validation," a shift that aligns generative assessment with design culture's concern for meaning, context, and interpretive accountability.

A two-layer evaluation logic provides a practical solution: it makes ECF divergence visible and it prevents high technical scores from functioning as implicit proof of cultural fit. This separation supports clearer claims in research reporting, because it allows authors to state precisely which form of performance improves.

Finally, the benchmark functions as a design instrument rather than as a post-hoc audit tool. In design research, instruments shape what becomes visible and therefore what becomes actionable. The cultural fidelity benchmark provides a repeatable way to locate failure modes, to compare systems under controlled prompt conditions, and to translate qualitative observations into structured intervention targets. Used iteratively, the benchmark supports prompt refinement, metadata redesign, and model adaptation decisions, and it provides a shared vocabulary for collaboration between designers, cultural experts, and technical teams. In this sense, benchmarking becomes part of the design process: a method for steering generative systems toward culturally accountable outputs through continuous evaluation and revision.

7.2 Implications for GLAM institutions and cultural policy

The results reposition GLAM institutions as active actors in the generative ecosystem. Museums and archives already function as validators of cultural knowledge through collection practices, cataloguing standards, and interpretive expertise. In the context of generative AI, this validating role extends into infrastructure provision: collections and their descriptive systems shape what becomes learnable, retrievable, and culturally legible in synthetic outputs. Cultural fidelity therefore depends on institutional decisions that historically belonged to heritage governance rather than to model development, especially as visual culture increasingly circulates through algorithmic infrastructures (Striphas 2015).

Metadata and access policies become especially consequential under this view. Cultural legibility in AI outputs draws from how objects, sites, and artworks are named, described, classified, and translated across languages. Controlled vocabularies, multilingual descriptors, provenance fields, and contextual notes supply epistemic density that supports cultural attribution and stylistic accuracy. Access policies shape which images circulate widely, which forms remain locally bounded, and which elements of cultural heritage become represented primarily through secondary, platform-driven iconographies. Crawford and Paglen's (2021) analysis of training images frames these selection effects as infrastructural, because dataset composition and labelling practices shape downstream representational capacity in AI systems. In practice, these policies influence whether AI systems learn heritage through high-quality documentation with strong provenance anchors or through fragmented, unevenly captioned web imagery shaped by ranking, categorisation, and platform governance (Gillespie 2018). Kuchtova (forthcoming) also frames institutional virtual archives as infrastructures that can strengthen democratic access and support resistance to censorship

through public availability and distribution, especially in contexts shaped by political pressure on cultural institutions.

Controlled collaboration offers a viable policy direction for supporting cultural sovereignty in low-resource contexts. This stakes a concrete governance role for GLAM institutions within the broader political economy of AI infrastructures and data extraction (Crawford 2021). Partnerships between GLAM institutions, universities, and technical teams can establish rights-cleared datasets, provenance documentation, and evaluation protocols that align model development with public cultural responsibilities. Such collaborations support accountability and reduce reliance on extractive pipelines that treat heritage collections as raw material for unregulated scraping, a dynamic widely discussed through the lens of data colonialism and large-scale appropriation (Couldry and Mejias 2019). They also create conditions for reciprocal benefit: institutions gain tools for education and interpretation, researchers gain structured cultural data, and communities gain representational agency through expert review and culturally situated quality gates. This approach treats cultural heritage as an infrastructural commons governed through consent, documentation, and shared evaluation, and it positions generative AI as a domain where cultural policy shapes the terms of visibility. This aligns with Krzykawski's framing of an East-Central European strategic choice around training data and cultural autonomy, which supports GLAM-led governance as a way to align cultural visibility with accountable infrastructures rather than default platform capture.

7.3 Implications for V4 toolmaking (text-to-video)

V4 toolmaking gains strategic value when it treats cultural fidelity as a shared regional infrastructure rather than as a country-by-country feature. Text-to-video applications amplify representational stakes through temporal continuity and narrative structure, and these properties call for common resources that support culturally situated generation across Hungarian, Polish, Czech, and Slovak contexts. A shared scenario library provides such a resource. It can assemble culturally grounded prompts and story fragments that encode regional diversity across domains, including architecture, folk traditions, and art history, while remaining comparable in structure for evaluation. In parallel, multilingual metadata functions as a cross-border interface layer: it connects local terms, diacritics, and domain vocabularies to aligned descriptors across languages and to an English mapping that supports tooling and interoperability. Cross-culture expert panels then provide the interpretive competence required for validation, ensuring that cultural cues remain legible within each context while supporting comparative diagnosis across the region.

Within this development ecology, the benchmark functions as a quality gate that links model iteration to cultural accountability. Ratings of cultural fit, stylistic accuracy, and technical quality provide structured criteria for release decisions, and the temporal extension for video supports stability checks across sequences.

A V4 text-to-video tool also functions as a testbed for culturally aware generative design. It supports comparative experimentation with data governance, metadata design, conditioning strategies, and interface guidance across multiple low-resource languages and cultural contexts. Workshops, public demonstrations, and educational deployments provide feedback loops that reveal how users interpret cultural references and how interface choices shape cultural outcomes. In this way, the tool becomes both a product and a research instrument: it operationalises cultural fidelity as a design goal, and it generates evidence about how culturally aware AI can support regional storytelling, education, and creative practice in ways that strengthen cultural visibility through accountable infrastructures.

8. CONCLUSION, LIMITATIONS, FUTURE RESEARCH

8.1 Conclusion

This article introduced epistemic cultural flattening (ECF) as a name for a structural gap in generative visual systems which explains how outputs can achieve strong technical plausibility while cultural provenance and stylistic accuracy remain unstable under culturally situated evaluation. The concept clarifies why polished images can still function as weak cultural evidence, especially in low-resource contexts where models resolve specificity through globally dominant templates. By framing this divergence through an epistemic interpretive framework (EIF), the paper positioned cultural fidelity as a design-relevant dimension of performance that shapes how images circulate as cultural references.

The paper also presented a cultural fidelity benchmark that makes this gap measurable across domains and models. By separating cultural fit, stylistic accuracy, and technical quality, the benchmark provides a repeatable instrument for comparative diagnosis. The accompanying typology translated benchmark divergence into a design vocabulary of failure modes, supporting interpretive clarity and actionable intervention targets.

Finally, the proposed V4 workflow demonstrated how measurement can inform intervention. The workflow treated cultural fidelity as an infrastructural design problem shaped by GLAM sourcing, multilingual metadata practices, controlled model adaptation, and iterative expert evaluation. In this framing, culturally aware text-to-video development becomes feasible through governance and evaluation structures that increase epistemic density, stabilise provenance anchors, and support culturally situated validation across Hungarian, Polish, Czech, and Slovak contexts.

8.2.1 Data availability

The study reports results from an ongoing benchmark corpus covering eight of ten targeted cultures at the time of writing. The published article presents aggregate findings due to space constraints. To improve transparency, the project makes available an online documentation package including the prompt inventory, scoring dimensions, codebook, and

benchmark summary tables at. To improve transparency, the project makes available an online documentation package—including the prompt inventory, scoring dimensions, codebook, and benchmark summary tables—which is linked in the online version of this paper. Release of the full image corpus remains conditional on permissions, platform terms, and publication constraints.

8.2 Limitations

This study operated under an English-prompt constraint that introduces translation and tokenisation bottlenecks for culturally specific terms, especially in low-resource language contexts. Inter-rater agreement remained moderate-to-low ($\alpha = 0.20–0.29$), consistent with the interpretive demands of culturally situated aesthetic evaluation. This level of agreement constrains claims to directional patterns and benchmark-level comparisons rather than fine-grained distinctions between individual cultures or prompts. Reference images functioned as grounding anchors and simultaneously reflected platformed visibility regimes. On the other hand their use requires careful attention to permissions in publication. The empirical focus treated Hungary as an anchor case within a comparative corpus. Results draw on eight of ten targeted cultures, evaluation of the remaining cultures was ongoing at the time of writing.

8.3 Future research

Future work can extend the typology and benchmark testing across the V4 region through shared scenario libraries and coordinated expert panels in Hungary, Poland, Czechia, and Slovakia. Text-to-video development calls for an explicit temporal cultural fidelity dimension that evaluates stability of cultural cues across sequences and narrative contexts. Controlled GLAM datasets and participatory metadata design offer an additional direction, enabling culturally grounded training pipelines that strengthen provenance anchors through multilingual descriptive systems. Work on bias loops in cultural heritage practice frames iterative mitigation through dataset governance, evaluation, and interpretive workflows, supporting this direction through an established practice-based model of intervention (Foka et al. 2025). User experience research can further clarify how diverse audiences interpret cultural cues in generated outputs, how interface choices steer cultural attribution, and how educational deployments shape trust, learning outcomes, and creative practice in culturally aware generative tools.

AUTHOR CONTRIBUTIONS

Brigitta Iványi-Bitter: conceptualisation; theoretical framing; writing (original draft); supervision. **Tibor Bacsi:** methodology; experimental setup; prompt strategy; image generation workflow; visual stimuli database curation. **Szilárd Szakács:** methodology; software; automated analysis pipeline; statistical modeling (linear mixed-effects models); validation; data visualisation. All authors: review and editing.

ACKNOWLEDGEMENTS

The corresponding author thanks the students of the course **Designing Against Bias in AI Systems** (Spring and Fall 2025) at MOME for contributions to test datasets, and feedback during the development of the benchmark. Special thanks go to **Vannawongpaisan Varissara, Luca Sára Kiss,** and **Zsófia Kérdy** for particularly insightful comments and support.

REFERENCES

Kuchta, Albín. 2026. "Digital Art Space as a Tool of Poietic Resistance or a Censorship Device." *Disegno: Journal of Design Culture*.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1–15.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Couldry, Nick, and Ulises A. Mejias. 2019. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford: Stanford University Press.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven: Yale University Press.

Crawford, Kate, and Trevor Paglen. 2021. "Excavating AI: The Politics of Images in Machine Learning Training Sets." Publication venue verification recommended for your final proof.

Foka, Anna, Gabriele Griffin, Diego Ortiz Pablo, Patrycja Rajkowska, and Sawsan Badri. 2025. "Tracing the Bias Loop: AI, Cultural Heritage and Bias-Mitigating in Practice." *AI & Society.*

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* New Haven: Yale University Press.

Hall, Stuart. 1997. *Representation: Cultural Representations and Signifying Practices.* London: Sage.

Janda, Jiří. 2026. "The Liminality of Generative Creation." *Disegno: Journal of Design Culture.*

Keszeg, Anna. 2026. "The Paprika-Effect." *Disegno: Journal of Design Culture*.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology.* 2nd ed. Thousand Oaks, CA: Sage.

Krzykawski, Michał. 2026. "Beyond Computational Illusion: Futures Worth Wanting for Artistic Practices and Technical Cultures." *Disegno: Journal of Design Culture*.

Malinowska, Ania. 2026. "AI Assimilationism: The Cultural Flattening of Localities in Generative Models."" *Disegno: Journal of Design Culture*.

Manovich, Lev. 2019. *AI Aesthetics.* Moscow: Strelka Press.

Markova, Katerina. 2026. "Against Collective Vulnerability: Understanding Cultural Alignment in LLMs." *Disegno: Journal of Design Culture*.

Striphas, Ted. 2015. "Algorithmic Culture." *European Journal of Cultural Studies* 18 (4–5): 395–412.