applied sciences

MDPI

Article

A Trustworthiness Assessment Method for AIGC-Based Animation Translation of Traditional Narrative Images in Architectural Decoration

Zihan Zhu [1,†], Xianying Fan [1,†], Sihui Chen [1], Zhihao Yu [1], Shunhe Chen [1,*] and Kaida Chen [2,*]

[1] College of Landscape Architecture and Art, Fujian Agriculture and Forestry University, No. 63 Xiyuangong Road, Minhou County, Fuzhou 350100, China; 12519108022@fafu.edu.cn (Z.Z.); 000q211006@fafu.edu.cn (X.F.); 52519053015@fafu.edu.cn (S.C.); 52519053027@fafu.edu.cn (Z.Y.)

[2] College of Fine Arts and Design, Jimei University, Xiamen 361021, China

* Correspondence: 000q151015@fafu.edu.cn (S.C.); kaidachenwhu@gmail.com (K.C.)

† These authors contributed equally to this work.

Abstract

Artificial intelligence-generated content (AIGC) image-to-video systems have lowered the technical threshold for presenting traditional cultural content in motion, while generated motion may alter characters, props, narrative relationships, craft features, composition, and symbols. This study implements a trustworthiness assessment workflow for traditional narrative images in architectural decoration, linking cultural evidence dossiers, sample-adapted prompts, computer-vision measurements, and blinded human coding. Eight source images were used to generate 120 videos under three prompt conditions, with five outputs per image–condition cell. Four coders assessed the videos, five experts reviewed the coding instrument, and two new coders independently assessed 48 videos. Primary deviation rates were 7.12%, 7.37%, and 9.07% for conditions A, B, and C, respectively, with no significant blocked difference after correction (Friedman $\chi^2(2) = 1.400$; Holm-adjusted $p = 1.000$). Mean item-level Krippendorff's alpha was 0.250 across 14 estimable items; the independent-review composite ICC (A,1) was 0.402. Additional analyses examined coder sensitivity and the grouped generation structure. The original association involving inter-frame stability SSIM was not supported in within-image–condition comparisons ($\rho = -0.052$; permutation $p = 0.610$). An exploratory association was observed between source-image SSIM and human-coded deviations within these cells ($\rho = -0.366$; Holm-adjusted permutation $p = 0.021$). The study provides a documented procedure for comparing generated outputs and identifies limits in prompt effectiveness and the reproducibility of human judgments.

Keywords: traditional narrative images; architectural decoration; AIGC; image-to-video; animation translation; trustworthiness assessment; cultural-heritage dissemination

Check for updates

Academic Editor: Dong Hyun Jeong

Received: 22 September 2026
Revised: 19 October 2026
Accepted: 23 October 2026
Published: 25 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Image-to-video generation is increasingly being introduced into the digital presentation of cultural heritage, shifting part of the production process from manually created animation to low-cost generative workflows. Current applications often emphasize whether a static heritage image can be made to move, while giving less attention to whether the generated motion preserves the cultural information encoded in characters, props, narrative

relationships, composition, and craft form. This issue is especially relevant to narrative images embedded in traditional architectural decoration, where literary subject matter, spatial organization, and material craftsmanship jointly determine the meaning of the image. The present study therefore treats AIGC-based animation translation as a pre-dissemination assessment problem and examines how deviations introduced during image-to-video generation can be identified, quantified, and reviewed.

The digitization of cultural heritage has expanded from documentation and virtual display to generative content production, making the reliability of cultural information a central object of assessment. The Nara Document on Authenticity incorporates form, materials, traditional techniques, use, and cultural context into judgments of authenticity [1]. Heritage interpretation must also be grounded in verifiable sources, maintain interpretive transparency, and respect cultural context [2]. Denard further emphasizes documenting evidence sources, production processes, and interpretive decisions so that digital outputs remain traceable and verifiable [3]. Research on authenticity further shows that authenticity judgments are shaped by cultural values and social context [4]. Rahaman conceptualizes digital heritage interpretation as an interaction among evidence, media, audiences, and interpretive practice [5], and translates principles of real-world heritage interpretation into requirements for evidence organization, audience understanding, and feedback in digital media [6]. Long-term reuse of digital resources further depends on traceable provenance, intended use, and update mechanisms [7]. Together, these studies establish an evidence-, process-, and context-based foundation for assessing generated content.

Immersive technologies have broadened the forms of cultural-heritage presentation and have also encouraged content quality and interpretive effectiveness to be assessed separately [8]. Requirements for accuracy, completeness, and error documentation in photogrammetry and three-dimensional scanning provide a reference for analyzing errors in generated dynamic content [9]. As artificial intelligence has entered the production and dissemination of cultural-heritage content, concerns have intensified regarding transparency, accountability, and risks to cultural context [10]. Generative models may inherit biases from their training data, which can distort cultural representation [11]. Such biases may accumulate across collection, digitization, training, generation, and dissemination [12]. Research on trustworthy artificial intelligence therefore emphasizes explainability, bias mitigation, and human oversight [13]. A recent assessment protocol has combined standardized generation, expert evaluation, and automated checks to examine AI-generated cultural-heritage content [14].

Image-to-video generation adds a temporal dimension to static generation and introduces cross-frame identity drift, motion substitution, structural deformation, and changes in event relationships. Traditional narrative images organize stories through characters, objects, space, and gesture [15], while architectural decoration also uses these visual relationships to convey narrative themes [16]. Video diffusion models probabilistically generate spatial and temporal information jointly, meaning that identical inputs may still produce different outputs [17]. Prompt wording was held constant within each condition because prompt components are controllable inputs in generation [18]. Five repeated outputs were generated for each image–condition cell to characterize stochastic variation under otherwise matched inputs.

SSIM converts changes in image structure into reproducible numerical measurements [19], whereas character identity, key props, and narrative events are better assessed through human coding and agreement analysis [20]. Accordingly, this study integrates cross-frame visual stability, preservation of narrative relationships, and shifts in cultural information within a single assessment workflow.

Building on this research foundation and the identified problem, the study addresses three research questions:

RQ1: What observable types of deviation occur when AIGC image-to-video systems animate traditional narrative images in architectural decoration?
RQ2: Do the types and extent of deviations change after narrative-evidence constraints and full-evidence constraints are introduced?
RQ3: How can computer-vision measurements and culturally informed human review be combined in a traceable assessment workflow, and how consistent are the resulting judgments and associations?

The study positions AIGC-based animation translation as a pre-dissemination assessment task. It evaluates an evidence-informed workflow linking cultural-evidence dossiers, sample-adapted prompts, computer-vision measurements, and blinded human coding across eight architectural-decoration images and 120 videos. Its contribution lies in defining video-level cultural deviations, documenting repeated generation, and examining condition effects and measurement consistency within a source-image-blocked design. Coded judgments remain connected to their source images and evidence references, while disagreement is retained for reliability and sensitivity analyses.

2. Related Work

2.1. Cultural-Heritage Authenticity and Digital Interpretation

Authenticity assessment has expanded from material authenticity to provenance, context, and value judgments; digital representation therefore requires explicit evidence types and interpretive foundations [4]. Digital heritage interpretation frameworks conceptualize data acquisition, content construction, dissemination, and feedback as a continuous process [5]. On this basis, the present study conducts trustworthiness assessment after content generation and before dissemination.

Sustainable digital heritage resources require metadata that preserve outputs, source materials, production processes, and scholarly uses [7]. Accordingly, this study freezes the source images, evidence dossiers, prompts, videos, blinded-assessment mappings, and statistical files. Research on artificial intelligence ethics also calls for the assessment of technical accuracy, cultural responsibility, contextual plausibility, and potential risk [10].

Galindo-Durán et al. [14] propose a protocol for image-based heritage content that combines standardized generation, expert evaluation, automated multimodal checks, and an analytical plan. The present study applies evidence-informed assessment to completed image-to-video experiments. It evaluates temporal changes, repeated outputs, source-image-blocked condition comparisons, and sensitivity to coder disagreement. These contributions concern implementation and testing of a video assessment procedure.

2.2. Visual Information Structure in Traditional Narrative Images

Narrative images organize time, events, and viewing direction within static pictures [21]. Animation reorganizes these cues, so assessment must address character appearance and action, interpersonal relationships, causal connections, and viewing sequence [22].

Viewers infer actions and events in static images from cues such as gesture and spatial relationships [23], meaning that even slight generated motion may change narrative interpretation. Static narratives also depend on spatial configuration and symbolic objects [24]. Human assessment therefore codes characters, key props, spatial relationships, and overall composition together.

Narrative images follow culturally specific organizational conventions. Chinese visual narratives have developed distinctive forms of composition, text–image interaction, and character relations [25], while temple decoration translates literary stories through

combinations of figures, objects, and scenes [16]. Images of historical stories also employ modes such as momentary narration and layered continuous narration [26]. Animation assessment must therefore be grounded in the evidential structure of each image.

2.3. Image-to-Video Generation and Prompt Constraints

Video diffusion models improve generation quality through latent-space modeling and temporal modules, but temporal consistency continues to constrain output stability [27]. Training-data selection, camera-motion modeling, and conditional control all affect the output [28]. Cascaded models must coordinate semantic accuracy, visual detail, and spatiotemporal continuity across generated frames [29].

Open-domain image animation emphasizes preservation of the input image's appearance, yet motion generation may still cause detail loss or displacement [30]. Temporal residual learning can strengthen alignment between subsequent frames and the input image [31]. Accordingly, this study separately measures similarity between the source image and generated frames and the cross-frame stability of characters, props, composition, and narrative relationships.

Prompts are an important control variable in generation. Prompt modifiers can be classified into subject, style, quality, constraint, and other types, providing an operational basis for progressively adding narrative, visual, craft, and exclusion information [32]. Research on diffusion-model outputs has measured bias as systematic differences in social representation [33], while imbalances and harmful patterns in multimodal datasets show that training data can shape model representations [34]. Drawing on this literature, the present study operationally defines deviation as changes in characters, props, structure, and event relationships, and tests prompt-constraint effects through between-condition comparisons.

2.4. Computer-Vision Assessment and Human Coding

PSNR measures pixel-level error, whereas SSIM measures local luminance, contrast, and structural relationships; together, they provide complementary descriptions of visual change [35]. DINOv2 learns visual representations without supervision from a specific downstream task [36]. This study uses cosine similarity between DINOv2 features to quantify high-level visual similarity between source images and video frames. Machine metrics quantify changes in pixels, structure, and visual features, while human review evaluates character identity, narrative relationships, and cultural context.

Video assessment must also measure motion and cross-frame stability. The Farneback method estimates motion fields between adjacent frames [37], although optical flow is also affected by texture, boundaries, occlusion, and non-rigid motion [38]. The FVD framework compares distributions of spatiotemporal video features, showing why video evaluation must account for temporal information rather than isolated frames alone [39]. The present study therefore combines measures of pixel error, structural preservation, high-level visual features, cross-frame stability, and motion magnitude.

Human coding is used to identify characters, props, structures, and narrative events. Content analysis requires explicit units of analysis, categories, and operational definitions, which provide the basis for the D01–D15 binary items [40]. Agreement indices must match the number of coders and the measurement scale [41], while coder training, blinded assessment, and documentation of disagreement improve process transparency [42]. This study uses Krippendorff's alpha as the principal chance-corrected index of item-level agreement [43]. Rare binary events may yield high observed agreement alongside low kappa values [44]. Event prevalence, observed agreement, and chance-corrected coefficients are therefore reported together.

3. Methods

3.1. Research Objective and Overall Workflow

This study defines the trustworthiness of AIGC-based animation translation as the degree of correspondence between a generated video and documented textual, visual, and craft references. Deviations encompass characters and props, narrative relationships, composition, craft and symbolic information, motion appropriateness, and overall misleading risk.

The study proceeded through evidence extraction, prompt grouping, repeated generation, computer-vision assessment, blinded human assessment, statistical comparison, and deviation reporting. Retained source files, prompt texts, assessment mappings, and analysis records provide the basis for tracing a coded result to its input and generation condition [3]. Figure 1 summarizes the workflow.

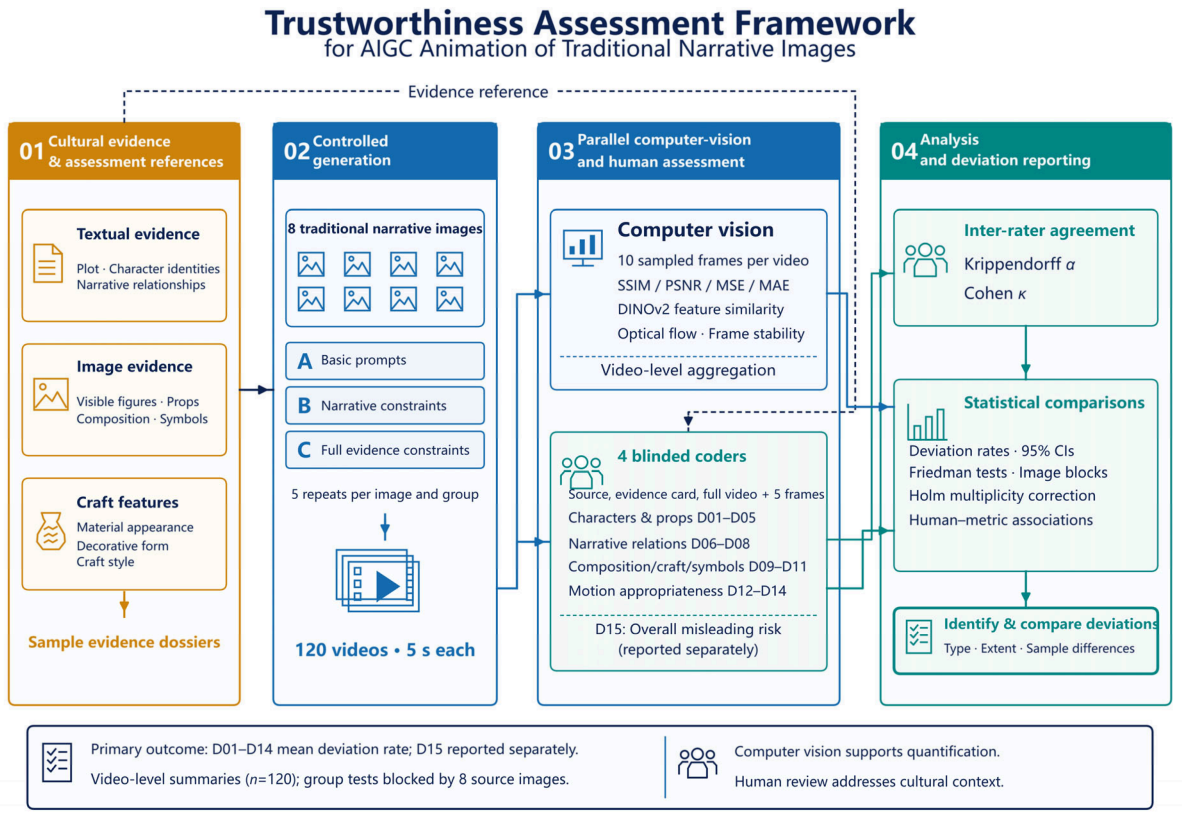

Figure 1. Trustworthiness assessment framework for AIGC-based animation translation of traditional narrative images.

The following subsections operationalize these four linked stages: evidence definition, controlled generation, parallel assessment, and deviation reporting.

3.2. Sample Selection and Cultural Evidence Dossiers

This study focused on hand-crafted narrative images integrated into architectural decoration. The sample was not intended to represent handscroll paintings, New Year prints, or book illustrations. Images were assembled purposively to vary narrative subject and craft medium. Selection prioritized an identifiable narrative subject, discernible principal figures and props, and sufficient visual clarity and completeness for comparison with generated frames. The eight retained inputs constitute the corpus analyzed in this study. Selection criteria are summarized in Supplementary Section S1.

The final corpus comprised eight images, labeled F01–F08. Their subjects were The Monkey King Proclaims Himself Great Sage Equal to Heaven, The Eight Immortals Crossing

the Sea, Xue Rengui Rescues His Lord, Jiang Ziya Musters the Troops, Nine Generations Under One Roof, Five Sons Achieving Scholarly Success, Zhao Zilong's Single-Handed Rescue, and Guo Ziyi's Birthday Celebration. The sample spans porcelain-cut decoration, clay-relief decoration, murals, and gilt woodcarving. This purposive corpus supports analysis across the sampled narrative subjects and craft contexts; it does not estimate deviation prevalence across all traditional narrative images.

The source image served as the common visual reference for machine quantification and human review. Each evidence dossier separates the narrative reference, directly visible image elements, and craft features. These records identify relationships to preserve and interpretations or additions to exclude. The complete input image remained the reference for human review. A dossier template, a worked example for F04, and translated assessment cards are provided in Supplementary Section S5. The eight image samples and design are summarized in Table 1.

Table 1. Eight image samples and main experimental design.

Sample ID	Narrative Subject	Source Images	Condition A Videos	Condition B Videos	Condition C Videos	Total Videos	Frames/ Video	Machine-Analysis Frames
F01	Monkey King	1	5	5	5	15	10	150
F02	Eight Immortals	1	5	5	5	15	10	150
F03	Xue Rengui's rescue	1	5	5	5	15	10	150
F04	Jiang Ziya musters troops	1	5	5	5	15	10	150
F05	Nine generations	1	5	5	5	15	10	150
F06	Five sons	1	5	5	5	15	10	150
F07	Zhao Zilong's rescue	1	5	5	5	15	10	150
F08	Guo Ziyi's birthday	1	5	5	5	15	10	150
Total	—	8	40	40	40	120	10	1200

Note: Statistical inference used the eight source images as blocks. The 120 videos were repeated generation outputs, not 120 independent cultural-heritage objects.

3.3. Prompt Conditions and Controlled Generation

Three sample-adapted prompt conditions were compared. All conditions requested a 5 s video, preservation of the input image's complete composition and aspect ratio, subtle motion, no further cropping or stretching, and no readable text. Condition A provided basic animation instructions and sample-specific preservation requirements. Condition B specified the narrative situation, character roles, and relationships to be expressed through motion. Condition C provided more detailed instructions concerning visible figures and props, spatial relationships, craft appearance, and prohibited additions or narrative substitutions. The comparisons assess the complete prompt packages used for each sample.

Prompt construction drew on the textual, visual, and craft components of the evidence dossiers. Narrative references supplied the scene and character relationships; visible-image observations supplied figures, props, positions, and compositional relationships; craft observations supplied material and decorative requirements. These components became image-specific instructions. For F04, the narrative specified a commander issuing orders and generals responding, while visual and craft instructions preserved weapons, figure arrangement, dragon ornament, and roof-ridge appearance. The 24 original Chinese prompts, their English translations, and a shared-versus-specific component matrix are supplied in Supplementary Section S2. Prompts followed principles of keyword organization and constraint reuse [18,32].

Five retained outputs were generated for each image under each prompt condition, giving 120 videos. The authors used Seedance 2.0. The requested duration was 5 s; retained files contained 121 frames at 24 fps, corresponding to approximately 5.04 s. Output dimensions varied by sample and are documented in Supplementary Section S3 and the accompanying metadata table. Archived texts record requested generation conditions.

The F01 A and B prompts allowed slow camera movement, whereas the remaining archived prompts emphasized camera stability. Several A prompts already contained sample-specific preservation requirements. The design combines sample-specific information with the A/B/C comparison and does not identify the independent effect of each prompt component.

Videos were assigned randomized identifiers V001–V120. These identifiers concealed prompt condition and repetition number. Coders received the corresponding source image and evidence card as assessment references. Figure 2 summarizes the eight inputs and balanced 8 × 3 × 5 design.

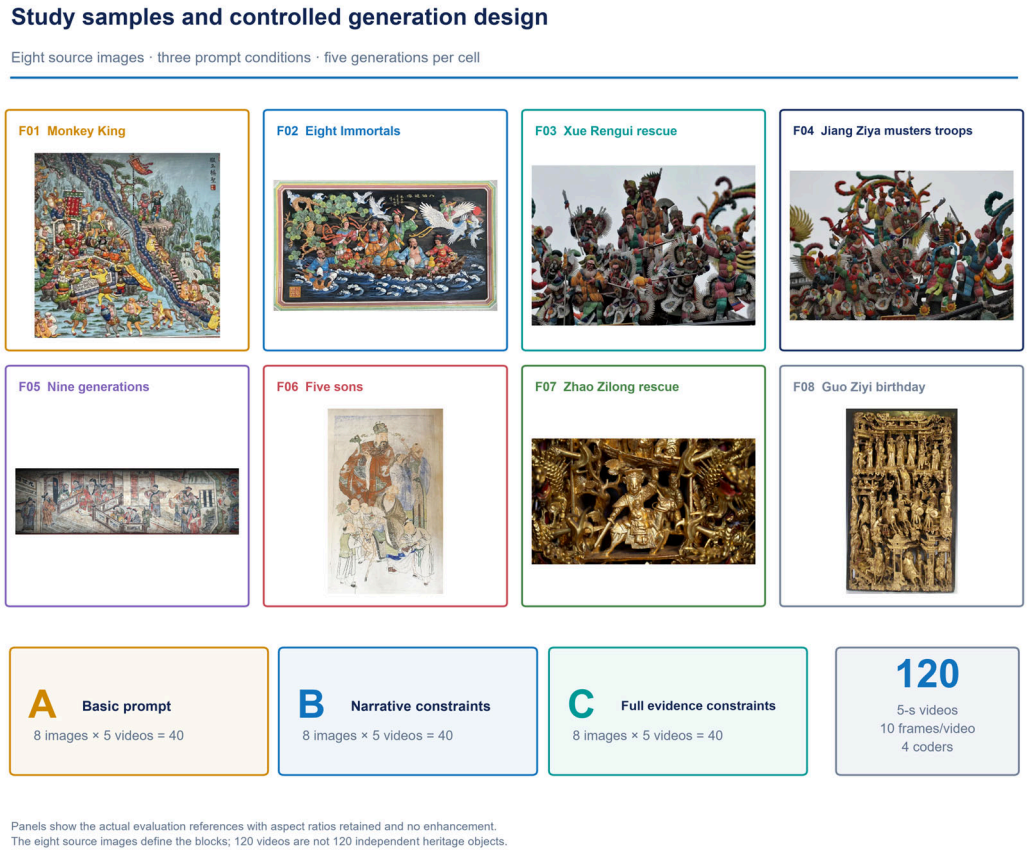

Figure 2. Image samples, prompt conditions, and repeated-generation design.

This balanced 8 × 3 × 5 design gives every source image equal representation in each prompt condition and defines the video-level dataset used in the subsequent assessment.

3.4. Sampled Frames and Computer-Vision Metrics

Computer-vision analysis used the video as the observational unit. Ten frames were decoded at evenly spaced requested times from 0 to 4.9 s, yielding 1200 analysis frames and nine adjacent-frame optical-flow intervals per video. Each source image was center-cropped to match the video aspect ratio and resized using OpenCV INTER_AREA interpolation. Human coders retained the complete input reference. This preprocessing excludes peripheral source content: for F05, approximately 69.1% of the original image area was retained

before resizing. Retained areas for all samples are reported in Supplementary Section S3 and the accompanying metadata table.

SSIM was calculated on grayscale float32 images using an 11 × 11 Gaussian window, σ = 1.5, an intensity range of 255, K_1 = 0.01, and K_2 = 0.03 [19]. MSE and MAE were averaged over RGB values, and PSNR used corresponding RGB MSE with a maximum intensity of 255 [35]. Source-image metrics were averaged across ten sampled frames. DINOv2 features were extracted using facebook/dinov2-small without fine-tuning [36]. The saved image processor resized the shortest edge to 256 pixels, applied an additional 224 × 224 center crop, and normalized RGB values with the model's stored configuration. Cosine similarity was calculated between L_2-normalized 384-dimensional CLS vectors. The source-image matching crop and the model's subsequent crop define the regions represented by this metric. SSIM and DINOv2 definitions are given in Equations (1) and (2):

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{1}$$

$$S_{\mathrm{DINO}} = \frac{f_o^T f_t}{\|f_o\|_2 \|f_t\|_2} \tag{2}$$

Inter-frame stability SSIM compared each of the nine later sampled frames with the first sampled frame. Farneback flow was calculated between consecutive sampled frames after proportional resizing to a width of 640 pixels [37]. Parameters were pyr_scale = 0.5, levels = 3, winsize = 15, iterations = 3, poly_n = 5, poly_sigma = 1.2, and flags = 0. Mean flow magnitude was averaged across nine intervals and represents pixel displacement per sampled interval in resized images [38]. The full model snapshot, normalization, numerical safeguards, and processing configuration are provided in Supplementary Section S3 and the configuration file. Inter-frame stability and mean optical flow are defined in Equations (3) and (4):

$$S_{\mathrm{stab}} = \frac{1}{T-1}\sum_{t=2}^{T} \mathrm{SSIM}(F_1, F_t) \tag{3}$$

$$\overline{v} = \frac{1}{(T-1)HW}\sum_{t=1}^{T-1}\sum_{i=1}^{H}\sum_{j=1}^{W}\sqrt{u_{tij}^2 + v_{tij}^2} \tag{4}$$

Frame-level measures were aggregated within each video, producing 120 video-level observations. Machine metrics describe pixel differences, structure, high-level visual features, and motion, whereas human review evaluates characters, narrative content, cultural context, and dissemination risk.

3.5. Blinded Human Assessment and Coding Framework

Four principal coders independently evaluated all 120 videos and completed D01–D15. Coding materials comprised the complete generated video, the corresponding input image, five review keyframes extracted at 0, 1, 2, 3, and 4 s, the sample evidence card, and the coding manual. The complete video was the unit of analysis, and each item was assigned once per video. Item definitions and assessment instructions appear in Supplementary Section S4.

The four principal coders were students in design-related programs. They read the sample evidence cards and reported understanding the represented narratives. No formal coder-training program was conducted. The cards supplied a common reference for narrative relationships, visible elements, craft features, and excluded interpretations. Coders were instructed to base judgments on these materials, avoid adding information from personal memory, and use the indeterminate category when evidence was insufficient.

They evaluated videos independently, and their original judgments were retained without consensus adjudication. Item-level coding agreement and deviation prevalence are reported in Table 2.

Table 2. Coding agreement and deviation prevalence for D01–D15.

Item	Krippendorff α	Complete Videos, *n*	Four-Coder Complete Agreement	Mean Pairwise Agreement	Deviation Prevalence	Valid Ratings, *n*	Indeterminate, *n*
D01	0.154	120	95.0%	97.2%	1.67%	480	0
D02	0.462	118	94.1%	96.9%	2.93%	478	2
D03	0.232	120	89.2%	94.4%	3.75%	480	0
D04	0.164	116	56.9%	76.2%	17.02%	476	4
D05	0.435	120	70.0%	83.2%	18.13%	480	0
D06	0.183	119	89.9%	94.7%	3.34%	479	1
D07	0.347	120	90.0%	94.0%	4.79%	480	0
D08	NA	120	100.0%	100.0%	0.00%	480	0
D09	0.493	120	96.7%	98.3%	1.67%	480	0
D10	0.058	120	93.3%	96.5%	1.88%	480	0
D11	−0.002	120	98.3%	99.2%	0.42%	480	0
D12	0.534	119	68.9%	82.3%	25.68%	479	1
D13	0.301	119	79.0%	88.3%	9.19%	479	1
D14	0.120	120	50.0%	72.2%	19.58%	480	0
D15	0.022	116	68.1%	83.8%	9.03%	476	4

Note: NA = not estimable. D08 contained only zero judgments; its chance-corrected alpha was not estimable. Mean alpha was 0.250 across the 14 estimable items. Observed agreement and deviation prevalence are reported alongside the coefficients.

3.6. Content Validity and Independent Reliability Review

After the main experiment, five experts rated D01–D15 on relevance, clarity, and operational feasibility using a four-point scale. Recorded fields spanned environmental design, traditional architectural decoration and digital dissemination; digital media and moving images; communication and design; craft research; and iconography. Content validity was measured using I-CVI, S-CVI/Ave, and S-CVI/UA [45], while Aiken's *V* measured standardized rating concentration [46]. This was a post hoc content-validity review, not pre-experiment item revision.

Two new coders who had not participated in the main coding independently reviewed 48 stratified-sample videos. The sample covered eight source images, three prompt conditions, and two videos from each image × condition cell. These coders were also students in design-related programs. They read the sample evidence cards and received no formal coder training. Item-level agreement was reported as observed agreement, Krippendorff's alpha, and Gwet's AC1 [47]. Video-level composite deviation rates were evaluated using ICC (A,1) [48] and Spearman correlation.

Content-validity and external reliability results are summarized in Table 3 and Supplementary Section S6.

3.7. Outcome Measures and Statistical Analysis

The primary human-coded outcome was the D01–D14 deviation rate for each video, pooling all valid zero/one judgments across the four coders; values coded as nine were excluded from the corresponding denominator. Each dimension score was the unweighted mean of its constituent item-level deviation rates, with each item rate calculated from

available coder judgments. D15 was analyzed separately as overall risk. In Equation (5), m is the valid-rating indicator, and values coded as nine are excluded from the denominator.

$$R_v = \frac{\sum_{c=1}^{C} \sum_{d=1}^{14} m_{vcd} y_{vcd}}{\sum_{c=1}^{C} \sum_{d=1}^{14} m_{vcd}} \quad (5)$$

Table 3. Content validity and external independent reliability validation.

A. Post Hoc Content-Validity Review of D01–D15 by Five Experts

Item	Dimension	Relevance I-CVI	Clarity I-CVI	Operational Feasibility I-CVI	Minimum k^*	Minimum V	Decision
D01	Characters and props	1.000	1.000	1.000	1.000	0.800	Retained
D02	Characters and props	1.000	1.000	1.000	1.000	0.800	Retained
D03	Characters and props	1.000	1.000	1.000	1.000	0.800	Retained
D04	Characters and props	1.000	1.000	1.000	1.000	0.800	Retained
D05	Characters and props	1.000	1.000	1.000	1.000	0.800	Retained
D06	Narrative relationships	1.000	1.000	1.000	1.000	0.933	Retained
D07	Narrative relationships	1.000	1.000	1.000	1.000	0.933	Retained
D08	Narrative relationships	1.000	1.000	1.000	1.000	0.800	Retained
D09	Compositional structure	1.000	1.000	1.000	1.000	0.800	Retained
D10	Craft style	1.000	0.800	0.800	0.763	0.733	Retained; refinement recommended
D11	Textual symbols	1.000	1.000	1.000	1.000	0.800	Retained
D12	Motion sufficiency	1.000	1.000	1.000	1.000	1.000	Retained
D13	Motion sufficiency	1.000	1.000	1.000	1.000	0.867	Retained
D14	Motion-induced disruption	1.000	1.000	1.000	1.000	0.867	Retained
D15	Dissemination risk	1.000	1.000	0.800	0.763	0.800	Retained; refinement recommended

B. External Independent Reliability Validation of 48 Stratified-Sample Videos by Two New Coders

Assessment Scope	Videos, n	Observed Agreement (95% CI)	Krippendorff α (95% CI)	Gwet AC1 (95% CI)	ICC (A,1) (95% CI)	Spearman ρ (95% CI)
D01–D14	48	93.45% (89.29–97.02%)	0.355 (0.238–0.499)	0.927 (0.875–0.969)	0.402 (0.101–0.745)	0.451 (0.218–0.648)
D15	48	77.08% (60.42–91.67%)	0.378 (−0.051–0.698)	0.639 (0.281–0.902)	—	—

(A) Note: k* denotes the modified kappa statistic, which adjusts I-CVI for chance agreement. "Minimum k*" is the lowest value across relevance, clarity, and operational feasibility for each item. S-CVI/Ave was 1.000 for relevance, 0.987 for clarity, and 0.973 for operational feasibility; overall S-CVI/Ave was 0.987. D10 and D15 were retained, with further refinement of their operational boundaries recommended on the basis of expert feedback; (B) note: confidence intervals use 30,000 bootstrap draws resampling the eight source images and retaining all sampled videos and both coders within each selected image. Thirteen D15 alpha draws were undefined; its interval uses the remaining 29,987 draws.

Between-condition comparisons used the source image as the block. For each source image, the five videos within each condition were first aggregated; Friedman tests were then performed across the eight blocks, with Kendall's W reported. The six human outcomes and seven computer-vision outcomes formed separate Holm-adjustment families [49,50]. The seven machine-human associations formed a further correction family within each specified correlation analysis. Proportions were rounded to 12 decimal places before blocked rank tests to preserve mathematical ties.

Confidence intervals for condition means were obtained from 30,000 percentile bootstrap draws of the eight source images. A/B/C summaries for each selected image were

retained together in every draw; paired contrasts used the same sampled images. The 2.5th and 97.5th percentiles defined the intervals, using seed 20261015 [51]. For the 48-video independent review, a source-image bootstrap retained the six sampled videos and both coders for each selected image together. This analysis used 30,000 draws and seed 20261018; undefined coefficient estimates were omitted from the corresponding interval.

The original association calculation first regressed the unranked human and machine variables on source-image and prompt-condition indicators, then calculated Spearman correlation [52] between residuals. Additional analyses examined grouped design and coder dependence. For the within-cell analysis, each variable was ranked across the 120 videos and its image × condition mean rank was subtracted. Centered ranks were correlated; significance was evaluated with 30,000 two-sided permutations restricted to the five repeats within each of 24 cells, using seed 20261016. Permutation p values were calculated as (exceedances + 1)/(30,000 + 1), followed by Holm correction over seven metrics. Leave-one-image, leave-one-coder, single-coder, and post hoc D04/D14-exclusion analyses were sensitivity analyses. The primary outcome was retained unchanged. In Equation (6), D denotes the video-level primary deviation rate, M_k denotes machine metric k, X contains an intercept and source-image and prompt-condition indicators, and R assigns average ranks for ties. Regression is performed before ranking the residuals. Supplementary Section S7 provides the analysis details and sensitivity results.

$$\rho_{res,k} = cor\left(R(D - X\beta_D),\ R\left(M_k - X\beta_{M_k}\right)\right) \quad (6)$$

Reliability of the principal coding was assessed primarily with Krippendorff's alpha for nominal data, together with pairwise Cohen's kappa, four-coder complete agreement, and mean pairwise agreement. Cohen's kappa is appropriate for nominal variables coded by two raters [20], whereas Krippendorff's alpha accommodates multiple coders and missing values [43].

4. Results

4.1. Data Completeness and Principal-Coding Agreement

All 120 videos were included in the combined human and machine analysis, with 40 videos in each of conditions A, B, and C and five repeated generations in every source image × prompt condition cell. Computer-vision processing was completed for all 1200 sampled frames. The four coders were expected to produce 7200 item–video judgments; 7187 were valid zero or one judgments, and 13 values coded as nine were treated as item-level missing data.

Across the 14 items with estimable coefficients, mean item-level Krippendorff's alpha was 0.250, ranging from −0.002 to 0.534. D08 contained 480 zero judgments and no positive judgments; its alpha was not estimable. Mean four-coder complete agreement across all 15 items was 82.6%, and mean pairwise agreement was 90.5%. Pooled pairwise positive agreement was 30.6% for D04, 29.1% for D14, and 10.9% for D15.

4.2. Content Validity and External Independent Reliability

The five experts completed 225 item–aspect ratings, of which 222 were scored three or four, yielding a qualifying rate of 98.7%. S-CVI/Ave values for relevance, clarity, and operational feasibility were 1.000, 0.987, and 0.973, respectively; overall S-CVI/Ave was 0.987, and S-CVI/UA was 0.933. All 45 I-CVI values were at least 0.78. Four experts endorsed the overall coverage, and none of the five recommended adding a first-order dimension. The lowest I-CVI was 0.800 for the clarity of D10 and the operational feasibility of D10 and D15, indicating that thresholds for craft style and overall risk require further specification.

In the 48-video independent review, pooled D01–D14 item–video agreement was 93.45% (source-image bootstrap 95% CI: 89.29–97.02%), Krippendorff's alpha was 0.355 (0.238–0.499), and Gwet's AC1 was 0.927 (0.875–0.969). For video-level composite rates, ICC(A,1) was 0.402 (0.101–0.745), and Spearman ρ was 0.451 (0.218–0.648). Pooled positive and negative agreement were 38.9% and 96.5%. For D15, observed agreement was 77.08%, alpha was 0.378, and AC1 was 0.639.

4.3. Prompt Conditions and Human-Coded Deviations

Primary deviation rates were 7.12% (95% CI: 4.67–10.34%), 7.37% (4.96–10.54%), and 9.07% (5.01–14.42%) for A, B, and C. The blocked Friedman test gave $\chi^2(2) = 1.400$, $p = 0.4966$, Holm-adjusted $p = 1.000$, and Kendall's $W = 0.0875$. The paired C−A difference was +1.95 percentage points (95% CI: −4.35 to +8.60).

In four leave-one-coder analyses, C−A remained positive, ranging from +1.47 to +2.24 percentage points; all blocked tests remained non-significant ($p = 0.4185$–0.6483). The four single-coder estimates also gave positive C−A differences, ranging from +1.07 to +3.39 percentage points. A post hoc analysis excluding D04 and D14 reduced C−A to +0.57 percentage points ($p = 0.4169$). The original primary outcome was retained in all sensitivity analyses.

Mean D15 values for conditions A, B, and C were 6.88%, 7.50%, and 12.50%, respectively. The Friedman test yielded $\chi^2(2) = 2.154$, $p = 0.3406$, Holm-adjusted $p = 1.000$, and Kendall's $W = 0.1346$, indicating no significant difference among conditions.

Motion-deviation rates were 23.47%, 13.54%, and 17.50% in A, B, and C. The Friedman test yielded $p = 0.0482$ and Holm-adjusted $p = 0.2889$ (Kendall's $W = 0.3792$). Figure 3 visualizes the human-coded deviation rates and 95% confidence intervals across conditions A, B, and C. Detailed estimates and blocked tests are reported in Table 4.

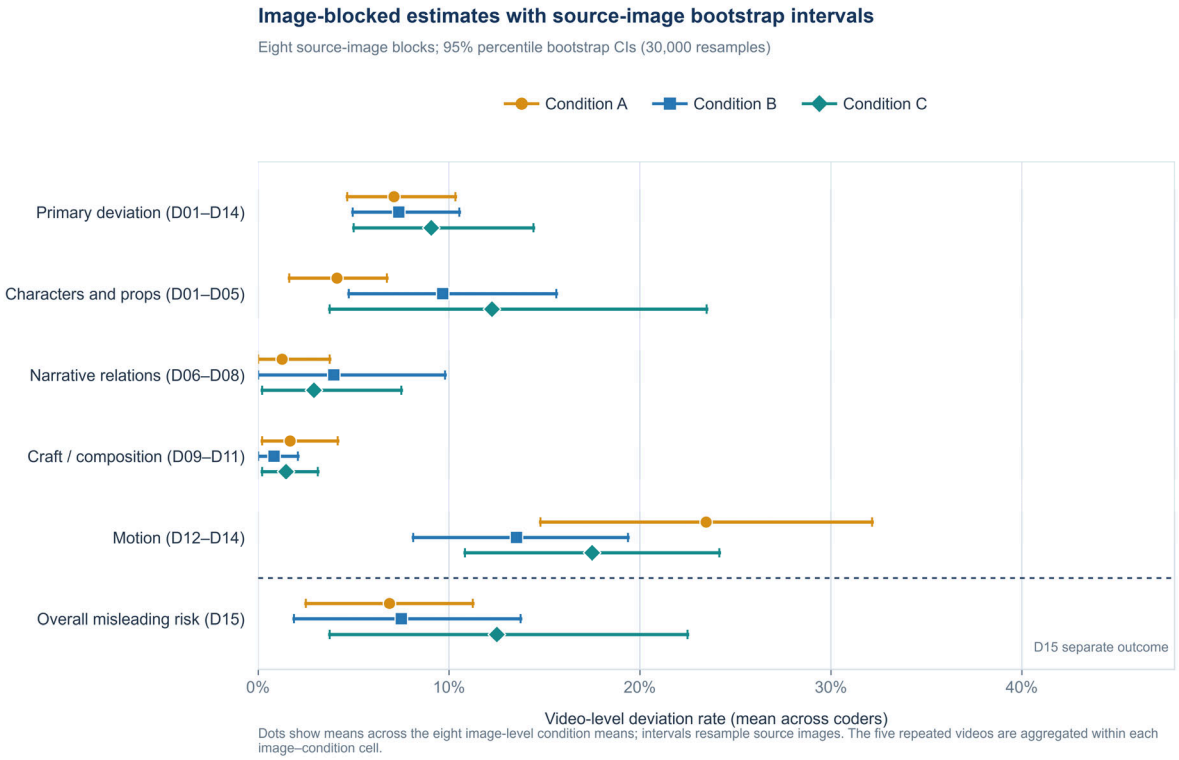

Figure 3. Human-coded deviation rates and 95% confidence intervals for conditions A, B, and C.

The paired C−A confidence interval includes zero, consistent with the blocked test that did not identify a condition difference.

Table 4. Human-coded deviation rates, Friedman tests, and Kendall's *W* across conditions A, B, and C.

Outcome	Condition A Mean (95% CI)	Condition B Mean (95% CI)	Condition C Mean (95% CI)	χ^2	Unadjusted *p*	Holm *p*	Kendall *W*
Primary deviation rate (D01–D14)	7.12% (4.67%, 10.34%)	7.37% (4.96%, 10.54%)	9.07% (5.01%, 14.42%)	1.400	0.4966	1.0000	0.0875
Characters and props (D01–D05)	4.13% (1.63%, 6.75%)	9.67% (4.75%, 15.63%)	12.25% (3.75%, 23.50%)	1.310	0.5194	1.0000	0.0819
Narrative relationships (D06–D08)	1.25% (0.00%, 3.75%)	3.96% (0.00%, 9.79%)	2.92% (0.21%, 7.50%)	1.000	0.6065	1.0000	0.0625
Composition, craft, and symbols (D09–D11)	1.67% (0.21%, 4.17%)	0.83% (0.00%, 2.08%)	1.46% (0.21%, 3.13%)	0.875	0.6456	1.0000	0.0547
Motion appropriateness (D12–D14)	23.47% (14.79%, 32.15%)	13.54% (8.13%, 19.38%)	17.50% (10.83%, 24.17%)	6.067	0.0482	0.2889	0.3792
Overall misleading risk (D15)	6.88% (2.50%, 11.25%)	7.50% (1.88%, 13.75%)	12.50% (3.75%, 22.50%)	2.154	0.3406	1.0000	0.1346

Note: Percentages are rounded to two decimals. Confidence intervals use 30,000 source-image bootstrap draws, retaining A/B/C together for each image. The eight source images are the inferential blocks. Holm-adjusted *p* values were calculated within six human outcomes; D15 is reported separately from D01–D14.

4.4. Dependence on the Source Image

The overall means concealed differences in direction across source images. In F01, the primary deviation rate decreased from 17.14% in A to 6.79% in B and 2.14% in C, whereas in F04 it increased from 5.00% to 17.14% and 25.71%. The direction and magnitude of descriptive condition differences varied across the eight source images.

After character, story, and exclusion constraints were added, motion direction became more explicit in some images, whereas complex multi-figure compositions showed more character redrawing and prop deformation. This pattern is consistent with the need for image-to-video models to coordinate semantics, detail, and spatiotemporal continuity simultaneously [29]. The source-image-specific patterns are shown in Figure 4.

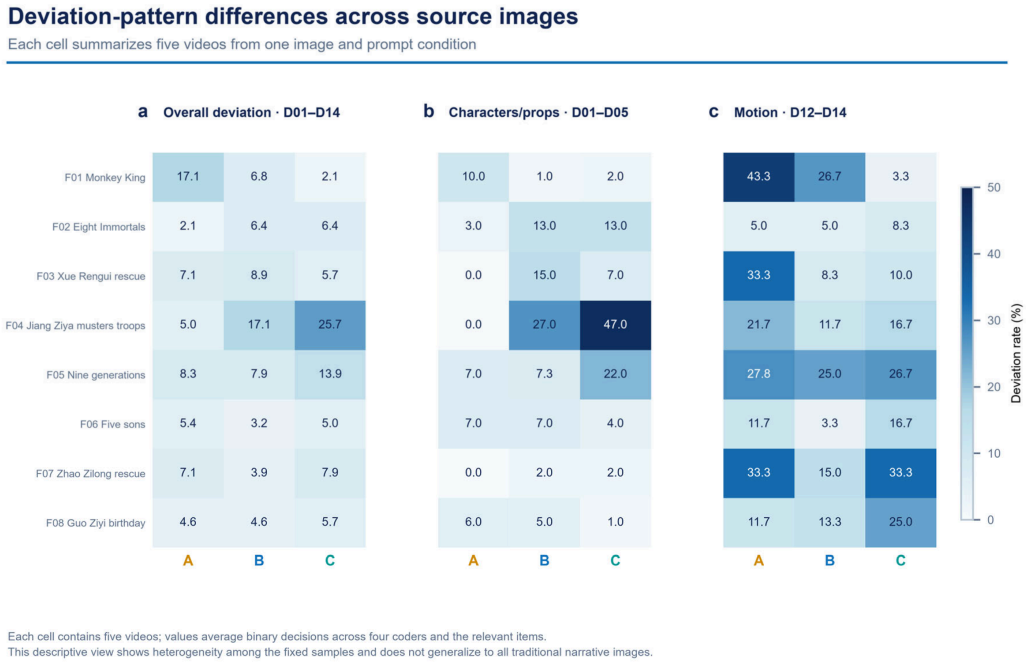

Figure 4. Source-image-specific deviation rates for the primary outcome, characters and props, and motion across conditions A, B, and C.

The figure therefore shifts interpretation from a single average prompt effect to source-image-specific response patterns.

4.5. Between-Condition Comparison of Computer-Vision Metrics

At the video level, mean source-image SSIM values for conditions A, B, and C were 0.414, 0.301, and 0.380; DINOv2 similarities were 0.938, 0.932, and 0.954; and mean optical-flow values were 2.414, 3.066, and 1.205. Between-condition inference used source-image-blocked tests to control for differences in sample composition, resolution, and craft.

Across the eight source-image blocks, SSIM, PSNR, MSE, and MAE each yielded an unadjusted p value of 0.0342 and a seven-metric Holm-adjusted p value of 0.2395. Adjusted p values for inter-frame stability SSIM, DINOv2 cosine similarity, and mean optical flow were all 0.5907. None of the seven metrics showed a significant between-condition difference after correction (Table 5A).

Table 5. Computer-vision condition effects and machine–human associations.

A. Effects of A/B/C Conditions (Friedman Blocked Test)

Outcome	Blocks	Friedman χ^2	Unadjusted p	Holm p	Kendall W
SSIM (vs. source)	8	6.750	0.0342	0.2395	0.4219
PSNR (vs. source)	8	6.750	0.0342	0.2395	0.4219
MSE (vs. source)	8	6.750	0.0342	0.2395	0.4219
MAE (vs. source)	8	6.750	0.0342	0.2395	0.4219
Inter-frame stability SSIM	8	1.750	0.4169	0.5907	0.1094
DINOv2 cosine similarity	8	3.250	0.1969	0.5907	0.2031
Mean optical flow	8	3.000	0.2231	0.5907	0.1875

B. Within-image × condition associations between machine metrics and the human-coded primary deviation rate (D01–D14)

Machine metric	n	Original residual ρ	Original Holm p	Within-cell rank ρ	Permutation p	Permutation Holm p	Decision
SSIM (vs. source)	120	−0.100	1.0000	−0.366	0.0039	0.0206	Significant after within-cell correction
PSNR (vs. source)	120	−0.098	1.0000	−0.335	0.0044	0.0206	Significant after within-cell correction
MSE (vs. source)	120	0.095	1.0000	0.350	0.0029	0.0205	Significant after within-cell correction
MAE (vs. source)	120	0.096	1.0000	0.352	0.0034	0.0206	Significant after within-cell correction
Inter-frame stability SSIM	120	−0.295	0.0074	−0.052	0.6102	0.6102	Not significant after within-cell correction
DINOv2 cosine similarity	120	−0.167	0.3442	−0.231	0.0294	0.0881	Not significant after within-cell correction
Mean optical flow	120	0.203	0.1568	0.162	0.1345	0.2691	Not significant after within-cell correction

Note: Panel A reports seven computer-vision outcomes; the human outcomes are reported in Table 4. Panel B shows the original residual coefficients and within-image × condition rank associations, tested by permutation among five repeats per cell (30,000 permutations). Holm correction was applied across seven machine metrics. Correlation does not imply causation.

4.6. Associations Between Machine Metrics and the Human-Coded Primary Deviation Rate

Reproduction of the original residual-Spearman procedure yielded an association between inter-frame stability SSIM and primary deviation rate of $\rho = -0.2953$ (conventional Holm-adjusted $p = 0.0074$). The additional analysis compared the five repeat videos within each image–condition cell. Under this specification, the inter-frame stability association was $\rho = -0.0518$ (permutation $p = 0.6102$; Holm-adjusted $p = 0.6102$).

Source-image SSIM showed a within-cell rank association of $\rho = -0.3662$ (permutation $p = 0.0039$; Holm-adjusted $p = 0.0206$). PSNR, MSE, and MAE showed associations in corresponding directions and remained significant within the seven-metric correction family. DINOv2 and mean optical flow did not remain significant after correction. Table 5B and Figure 5 report these sensitivity results alongside original residual coefficients.

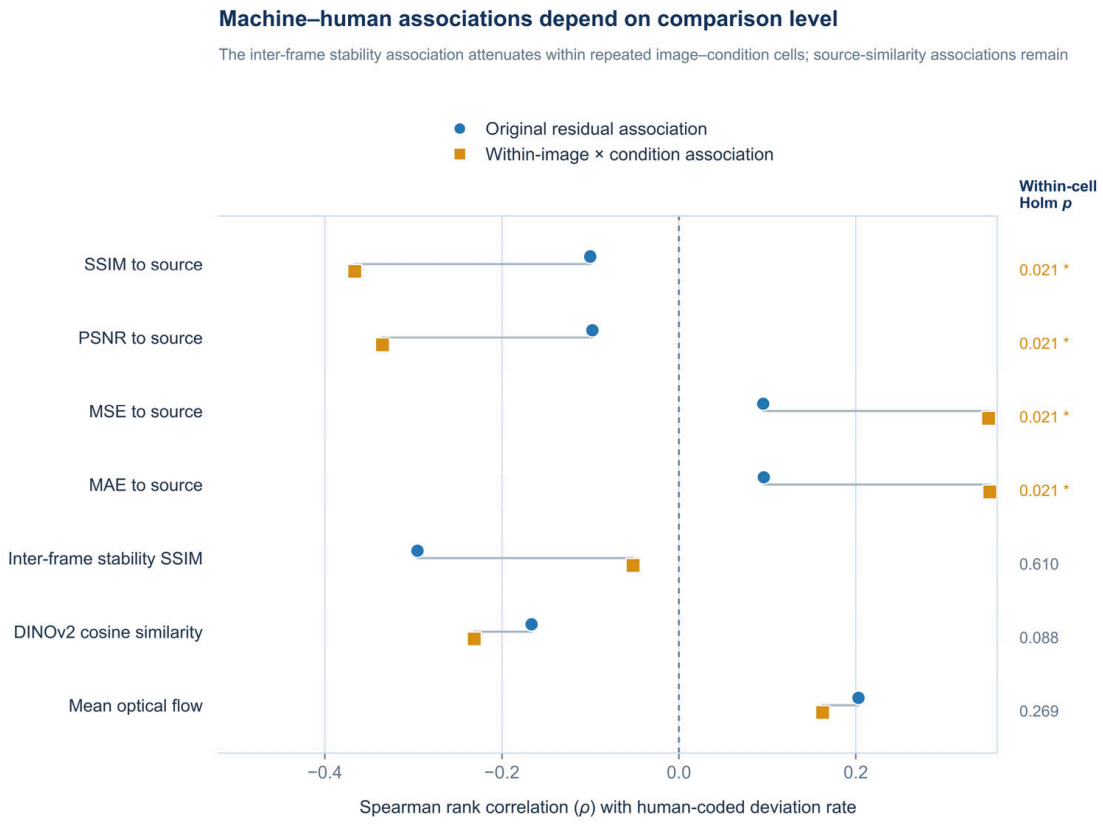

Figure 5. Original residual and within-image × condition associations between computer-vision metrics and human-coded deviations.

The source-image SSIM association remained negative when each source image was omitted in turn ($\rho = -0.420$ to -0.212). Removing one coder at a time yielded $\rho = -0.392$ to -0.314, with Holm-adjusted permutation p values of 0.0170–0.0285. Single-coder associations remained negative but did not remain significant after correction. Excluding D04 and D14 in a post hoc analysis gave $\rho = -0.3513$ and Holm-adjusted $p = 0.0758$. These results depend on metric, correlation specification, and judgments included in the outcome.

4.7. Visual Examples of Identified Deviations

Figure 6 illustrates three deviations coded by all four principal coders in videos generated from F04: absent internal motion in A3 (D12), figure deformation in B1 (D04), and key-prop change in C1 (D05). Source images and selected time points are paired to connect the coded decisions with their visual references. The examples illustrate the coding criteria and do not represent average condition performance.



Figure 6. Representative examples of unanimously coded deviations under prompt conditions A, B, and C. The full 5 s videos were the coding units; selected frames are shown for illustration.

5. Discussion

5.1. What the Assessment Workflow Establishes

This study links cultural evidence preparation, repeated generation, deviation measurement, and statistical comparison in a documented assessment procedure. Retained source files, prompt texts, coding references, and analysis records connect results to their inputs and generation conditions. Repeated outputs make variation within each image–condition cell observable, while coder sensitivity analyses expose variation in measurement. The workflow provides an implemented example of examining cultural evidence alongside temporal video measurements. Reliability results describe consistency achieved by participating coders, and machine associations describe relationships within this corpus.

Galindo-Durán et al. [14] propose an image-based protocol for cultural-heritage content. This study evaluates a completed image-to-video experiment by examining temporal changes, repeated outputs, source-image-blocked condition comparisons, and sensitivity to coder disagreement. Its contribution lies in implementing and testing a video assessment procedure.

The workflow links source evidence, prompt condition, generated video, video-level measurements, and human judgments. Comparisons among conditions and source images describe where deviations occurred in this corpus, while coder analyses show how measurement uncertainty affects those comparisons. These records support re-examination of reported outputs and judgments.

5.2. Why Evidence Constraints Did Not Consistently Reduce Overall Deviation

Condition C had the highest mean primary deviation rate, exceeding A by 1.95 percentage points. The paired interval extended from a reduction of 4.35 to an increase of 8.60 percentage points, so the data do not establish an advantage or a consistent deteriora-

tion. The most extensive prompt package did not yield a monotonic improvement in the retained outputs.

A possible explanation is competition among instructions concerning character relationships, object positions, motion, and craft appearance. The design cannot identify that mechanism because conditions alter several components together. The A prompts already contain some sample-specific preservation requirements, and camera instructions also vary across samples. The observed differences apply to these complete prompt packages. Greater prompt length should not be treated as a sufficient quality-control rule; repeated output assessment remains necessary.

Source-image characteristics and training-data representations may also shape generation outcomes. This experiment did not test these mechanisms. The direction and magnitude of descriptive condition differences varied across source images, motivating future experiments that measure image density, occlusion, and available motion regions directly.

These results redefine the role of prompt constraints within the assessment workflow. Textual, visual, and craft evidence specify what should be preserved, but they do not guarantee that the generated output will satisfy those requirements. Prompts should be treated as controlled experimental conditions whose actual effects are tested through repeated generation, dimension-specific coding, and image-level comparison. For complex traditional narrative images, assessment should focus on whether the generated result consistently preserves the characters, relationships, structures, and craft information specified by the evidence.

5.3. Complementarity Between Computer Vision and Human Review

The grouped sensitivity analysis changes the practical interpretation of SSIM. The original residual association involving inter-frame stability was not retained when comparisons were restricted to repeated outputs from the same image and prompt condition. Source-image SSIM showed a negative association with coded deviations under this restriction and remained supported in leave-one-coder analyses. Its association was weaker under single-coder and item-exclusion specifications. These findings support interpreting structural measurements alongside coding results within the sampled corpus.

SSIM measures structural similarity [19], DINOv2 represents high-level visual features [36], and optical flow describes image motion [38]. A video can remain structurally stable while lacking requested internal motion; legitimate animation can also reduce similarity to the input. Cultural assessment therefore requires the source image, narrative reference, and full video. The analyses do not establish a machine threshold for identifying culturally misleading outputs or locating problematic time intervals.

Trustworthiness in AIGC-based heritage animation cannot be fully represented by mechanical binary indicators. Computer vision quantifies structural and motion changes, while human review interprets those changes in context. Human judgments also depend on contextual knowledge and decision thresholds. Reporting evidence references, operational rules, and observed disagreement makes these dependencies visible.

5.4. Agreement and Subjectivity in Human Judgment

High observed agreement was dominated by agreement on absent deviations. In the principal coding, positive agreement was 30.6% for D04, 29.1% for D14, and 10.9% for D15. The independent review showed 38.9% positive and 96.5% negative agreement across D01–D14, with a composite ICC(A,1) of 0.402. Low event prevalence contributes to differences between observed and chance-corrected agreement [44], while weak positive-case agreement also reflects uncertainty in applying thresholds.

The between-condition conclusion was unchanged in leave-one-coder analyses, while the magnitude and statistical support of machine-human associations depended on coder inclusion and item selection. This separates the stability of the corpus-level comparison from reproducibility of individual judgments. The four-coder outcome provides the principal estimate, with variation across coders reported as measurement uncertainty [41,43].

Expert review supported relevance and coverage of the instrument [45,46]. Future applications should incorporate training with positive, negative, and boundary examples, craft-specific discussion, and calibration before independent coding [42]. Training should be evaluated prospectively, with remaining disagreement reported.

Expert ratings were lower for clarity of D10 and operational feasibility of D10 and D15. Future coding materials can add positive and negative examples for craft boundaries and specify the threshold for an overall misleading-risk judgment.

5.5. Implications for the Dynamic Dissemination of Cultural Heritage

Generative artificial intelligence lowers the cost of producing animated cultural-heritage content and moves quality control upstream to evidence preparation and pre-dissemination review. Research on artificial intelligence ethics emphasizes responsibility, transparency, and contextual risk [10], while research on bias loops shows that problems can accumulate across collection, digitization, training materials, generation, and dissemination [12]. Disseminating institutions should therefore document the source image, narrative basis, craft features, model version, prompt, and generation batch and retain the results of machine screening and human review. Linking each video to these evidence records enables later users to distinguish changes originating in the heritage source from those introduced by the generation model or by subsequent editing during dissemination.

In dissemination practice, the workflow provides a structure for retaining source reference, prompt condition, generated output, machine measurements, and human judgments together. Institutions can use this record to compare versions and document why further review or revision was requested. Automated prioritization, temporal localization, and acceptance or rejection rules are potential extensions. Their use would require independently validated thresholds, error rates, and decision criteria not estimated in this experiment.

5.6. Limitations and Future Research

The corpus contains eight purposively assembled images and five outputs per image–condition cell from Seedance 2.0. The sample supports analysis of these inputs and craft contexts. Wider generalization requires additional source images, cultural traditions, and generation models.

Prompt conditions combine narrative, visual, craft, and exclusion instructions, with some preservation requirements already present in A. The experiment estimates differences between the implemented packages. Factorial or ablation designs would be needed to identify individual component contributions.

Binary coding omits deviation severity and duration. Cultural interpretations were referenced to shared evidence cards but were not independently verified at the scene level. Limited agreement on positive cases and composite scores constrains the reproducibility of individual-video judgments. More detailed examples, craft-specific thresholds, and prospective calibration would strengthen evaluation of judgment reproducibility.

Spatial matching excludes peripheral source content, particularly in F05, and the DINOv2 processor adds a further center crop. Machine results describe the regions represented by the implemented preprocessing. Grouped correlation and coder sensitivity analyses were conducted during revision and remain exploratory. Future work can compare padding or region-based preprocessing and test screening thresholds on independent data.

Future research should expand the sample and number of repetitions, conduct cross-model validation and constraint-ablation studies, refine thresholds for craft style and overall risk, and test reproducibility on an independent dataset.

6. Conclusions

This study implemented a trustworthiness assessment workflow linking cultural evidence, repeated image-to-video generation, computer-vision measurements, and independent human judgments. Across eight source images and 120 videos, the three prompt conditions showed no significant difference in the primary deviation rate after correction. The most extensive prompt package did not consistently improve retained outputs. Additional analyses showed that the original association involving inter-frame stability SSIM depended on correlation specification. Source-image SSIM was associated with coded deviations in within-image–condition comparisons, while coder and item sensitivity analyses identified limits to that association.

The workflow provides a documented structure for comparing outputs and reporting disagreements. Its methodological contribution is the joint assessment of generation variation and measurement uncertainty. Further evaluation of coder training, cultural interpretation, and predictive thresholds is required before operational screening applications.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app162110562/s1, Supplementary File S1: Supplementary methods and supporting results (Sections S1–S7), original Chinese prompts with English translations, de-identified assessment data, computer-vision settings, configuration files, analysis scripts, and supporting tables.

Author Contributions: Conceptualization, Z.Z. and X.F.; methodology, Z.Z. and X.F.; software, Z.Y.; validation, X.F. and S.C. (Sihui Chen); formal analysis, Z.Z. and X.F.; investigation, Z.Z., X.F. and S.C. (Sihui Chen); resources, Z.Y.; data curation, S.C. (Sihui Chen); writing—original draft preparation, Z.Z. and X.F.; writing—review and editing, Z.Z., X.F., S.C. (Shunhe Chen) and K.C.; visualization, X.F. and Z.Y.; supervision, S.C. (Shunhe Chen) and K.C.; project administration, S.C. (Shunhe Chen). All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by Fujian Provincial Social Science Foundation Project (FJ2026C181).

Institutional Review Board Statement: Not applicable. This study involved the evaluation of AIGC-generated videos by coders and experts and did not collect personally identifiable, demographic, medical, sensitive, or private information.

Informed Consent Statement: The four principal coders, five content-validity experts, and two independent validation coders were informed of the study purpose and evaluation procedures and voluntarily agreed to participate. No personally identifiable, demographic, sensitive, or private information was collected or reported.

Data Availability Statement: The de-identified main and independent coding records, expert ratings, derived computer-vision measurements, original Chinese prompts and English translations, configuration files, analysis scripts, and numerical data underlying the reported tables and figures are provided in the Supplementary Materials. The full-resolution source images and generated video files are not included in the Supplementary Materials or publicly redistributed because reuse and redistribution permissions vary across the source images. Requests to inspect generated videos for scholarly verification may be directed to the corresponding author and will be considered where the applicable rights permit.

Acknowledgments: The authors thank the four principal coders, the five experts who participated in the post hoc content-validity review, and the two independent coders who completed the external reliability validation. The authors reviewed and validated all study materials, analyses, and outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. International Council on Monuments and Sites. *The Nara Document on Authenticity*; ICOMOS: Paris, France, 1994. Available online: https://www.icomos.org/en/charters-and-texts/179-articles-en-francais/ressources/charters-and-standards/386-the-nara-document-on-authenticity-1994 (accessed on 25 September 2026).
2. International Council on Monuments and Sites. *The ICOMOS Charter for the Interpretation and Presentation of Cultural Heritage Sites*; ICOMOS: Paris, France, 2008. Available online: https://www.icomos.org/images/DOCUMENTS/Charters/interpretation_e.pdf (accessed on 25 September 2026).
3. Denard, H. A new introduction to the London Charter. In *Paradata and Transparency in Virtual Heritage*; Bentkowska-Kafel, A., Baker, D., Denard, H., Eds.; Ashgate: Farnham, UK, 2012; pp. 57–71.
4. Stovel, H. Origins and influence of the Nara Document on Authenticity. *APT Bull.* **2008**, *39*, 9–17.
5. Rahaman, H. Digital heritage interpretation: A conceptual framework. *Digit. Creat.* **2018**, *29*, 208–234. [CrossRef]
6. Rahaman, H.; Tan, B.K. Digital heritage interpretation: Learning from the realm of real-world. *J. Interpret. Res.* **2017**, *22*, 53–64. [CrossRef]
7. Champion, E.M.; Rahaman, H. 3D digital heritage models as sustainable scholarly resources. *Sustainability* **2019**, *11*, 2425. [CrossRef]
8. Bekele, M.K.; Pierdicca, R.; Frontoni, E.; Malinverni, E.S.; Gain, J. A survey of augmented, virtual, and mixed reality for cultural heritage. *J. Comput. Cult. Herit.* **2018**, *11*, 1–36. [CrossRef]
9. Remondino, F. Heritage recording and 3D modeling with photogrammetry and 3D scanning. *Remote Sens.* **2011**, *3*, 1104–1138. [CrossRef]
10. Tiribelli, S.; Pansoni, S.; Frontoni, E.; Giovanola, B. Ethics of artificial intelligence for cultural heritage: Opportunities and challenges. *IEEE Trans. Technol. Soc.* **2024**, *5*, 293–305. [CrossRef]
11. Foka, A.; Griffin, G. AI, cultural heritage, and bias: Some key queries that arise from the use of GenAI. *Heritage* **2024**, *7*, 6125–6136. [CrossRef]
12. Foka, A.; Griffin, G.; Ortiz Pablo, D.; Rajkowska, P.; Badri, S. Tracing the bias loop: AI, cultural heritage and bias-mitigating in practice. *AI Soc.* **2025**, *40*, 5835–5847. [CrossRef]
13. Paolanti, M.; Frontoni, E.; Pierdicca, R. Towards trustworthy AI in cultural heritage. *npj Herit. Sci.* **2026**, *14*, 131. [CrossRef]
14. Galindo-Durán, A.; Prego de Oliver-López, C.; Bernal-Bravo, C. A methodological protocol for the generation and evaluation of AI-generated cultural heritage content. *J. Cult. Herit.* **2026**, *81*, 370–381. [CrossRef]
15. Murray, J.K. What is Chinese narrative illustration? *Art Bull.* **1998**, *80*, 602–615. [CrossRef]
16. Lee, H.L.; Tseng, C.H.; Liu, C.H. Narrative conveyance of story themes in temples. *Bull. Jpn. Soc. Sci. Des.* **2015**, *62*, 77–84. [CrossRef] [PubMed]
17. Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; Fleet, D.J. Video diffusion models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 8633–8646. [CrossRef]
18. Liu, V.; Chilton, L.B. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2022; pp. 1–23. [CrossRef]
19. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
20. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
21. Nanay, B. Narrative pictures. *J. Aesthet. Art Crit.* **2009**, *67*, 119–129. [CrossRef]
22. Fasnacht, H. The narrative characteristics of images. *Br. J. Aesthet.* **2023**, *63*, 1–23. [CrossRef]
23. Horváth, G. Visual imagination and the narrative image: Parallelisms between art history and neuroscience. *Cortex* **2018**, *105*, 144–154. [CrossRef] [PubMed]
24. Peterson, M. The production of narrative through static imagery: Examples from a peculiar medieval illustration. *Vis. Commun.* **2019**, *18*, 279–293. [CrossRef]
25. Qi, F.; Fauzi, T.A.; Yahaya, S.R. Representing tradition: The construction of culturally-specific visual narratives in Chinese picture books and hand scroll paintings. *J. Early Child. Lit.* **2025**, *25*, 607–636. [CrossRef]
26. Zhu, H. Two narrative modes and their formation mechanisms in the historical stories depicted in Han Dynasty pictorial arts: A study with a focus on The Orphan of the Zhao Family. *Theor. Stud. Lit. Art* **2024**, *44*, 58–68. (In Chinese)
27. Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S.W.; Fidler, S.; Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 22563–22575. [CrossRef]

28. Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv* **2023**, arXiv:2311.15127. [CrossRef]
29. Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; Zhou, J. I2VGen-XL: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv* **2023**, arXiv:2311.04145. [CrossRef]
30. Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; Wong, T.-T. DynamiCrafter: Animating open-domain images with video diffusion priors. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 399–417.
31. Zhang, Z.; Long, F.; Pan, Y.; Qiu, Z.; Yao, T.; Cao, Y.; Mei, T. TRIP: Temporal residual learning with image noise prior for image-to-video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Seattle, WA, USA, 2024; pp. 8671–8681.
32. Oppenlaender, J. A taxonomy of prompt modifiers for text-to-image generation. *Behav. Inf. Technol.* **2024**, *43*, 3763–3776. [CrossRef]
33. Luccioni, A.S.; Akiki, C.; Mitchell, M.; Jernite, Y. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2023.
34. Birhane, A.; Prabhu, V.U.; Kahembwe, E. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv* **2021**, arXiv:2110.01963. [CrossRef]
35. Horé, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In *Proceedings of the 20th International Conference on Pattern Recognition*; IEEE Computer Society: Los Alamitos, CA, USA, 2010; pp. 2366–2369. [CrossRef]
36. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.* 2024. Available online: https://openreview.net/forum?id=a68SUt6zFt (accessed on 22 October 2026).
37. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*; Springer: Berlin, Germany, 2003; pp. 363–370. [CrossRef] [PubMed]
38. Baker, S.; Scharstein, D.; Lewis, J.P.; Roth, S.; Black, M.J.; Szeliski, R. A database and evaluation methodology for optical flow. *Int. J. Comput. Vis.* **2011**, *92*, 1–31. [CrossRef]
39. Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. Towards accurate generative models of video: A new metric and challenges. *arXiv* **2018**, arXiv:1812.01717. [CrossRef]
40. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 4th ed.; Sage: Thousand Oaks, CA, USA, 2019.
41. Hallgren, K.A. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor. Quant. Methods Psychol.* **2012**, *8*, 23–34. [CrossRef] [PubMed]
42. O'Connor, C.; Joffe, H. Intercoder reliability in qualitative research: Debates and practical guidelines. *Int. J. Qual. Methods* **2020**, *19*, 1609406919899220. [CrossRef]
43. Hayes, A.F.; Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* **2007**, *1*, 77–89. [CrossRef]
44. Feinstein, A.R.; Cicchetti, D.V. High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [CrossRef] [PubMed]
45. Polit, D.F.; Beck, C.T. The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Res. Nurs. Health* **2006**, *29*, 489–497. [CrossRef] [PubMed]
46. Aiken, L.R. Three coefficients for analyzing the reliability and validity of ratings. *Educ. Psychol. Meas.* **1985**, *45*, 131–142. [CrossRef]
47. Gwet, K.L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29–48. [CrossRef] [PubMed]
48. McGraw, K.O.; Wong, S.P. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1996**, *1*, 30–46. [CrossRef]
49. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
50. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
51. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, NY, USA, 1993. [CrossRef]
52. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1904**, *15*, 72–101. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.