SAME STORY, ANOTHER CULTURE: NARRATIVE CULTURAL RE-INSTANTIATION FOR CROSS-CULTURAL VIDEO GENERATION

Anonymous authors
Paper under double-blind review

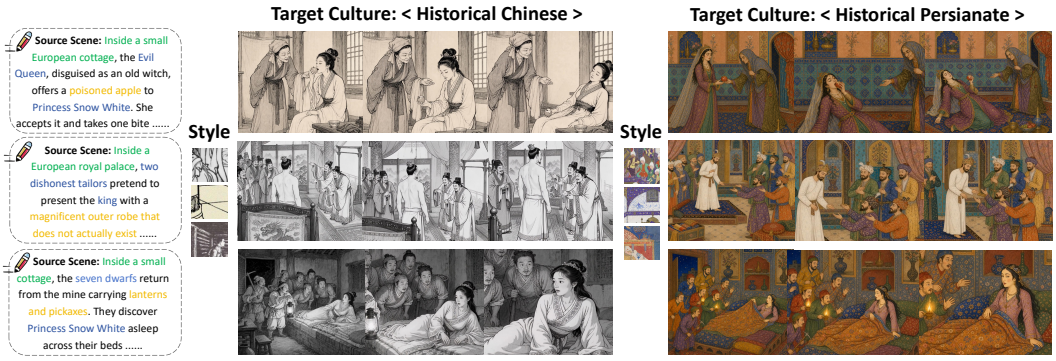

Figure 1: CultureWeaver re-instantiates source narrative scenes in different target-cultural settings using external cultural memories and frozen generative models.

ABSTRACT

Reinterpreting a narrative across cultures requires more than changing visual appearance: cultural substitutions must jointly preserve narrative functions and relations. We introduce *Narrative Cultural Re-instantiation*, the task of generating a target-cultural video from a source-cultural narrative while preserving its specified story structure. The central challenge is that individually plausible substitutions need not form a coherent whole. We propose **CultureWeaver**, a training-free framework that separates cultural correspondence from visual realization. An evidence-grounded *Visual Cultural Memory* supplies visually supported candidates annotated with narrative functions, affordances, and relations. CultureWeaver organizes retrieved candidates into a graph and jointly aligns them with the narrative through *relation-aware semi-relaxed structural transport*, balancing functional compatibility and relational consistency while allowing irrelevant candidates to receive negligible mass. Frozen generative models render the resulting cultural bindings into videos. We construct a benchmark of 200 narrative scenes across two target-cultural settings, with task-specific evaluation of narrative preservation, cultural fidelity, functional validity, and relational coherence. Automated evaluation and human judgments favor CultureWeaver over competing methods, while controlled ablations support the benefits of structured cultural memory and joint correspondence. Code and benchmark data are available in our anonymous repository.

1 INTRODUCTION

Stories have long crossed cultures by being retold rather than copied. For example, the Cinderella tale was recorded in ninth-century China as the story of Ye Xian, in which a lost shoe likewise leads a king to the heroine (Waley, 1947). Such retellings change characters, objects, and settings while preserving what Propp (2010) called narrative *functions*: acts defined by their role in the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

course of the action, independent of who performs them. Because their realization depends on culturally learned expectations (Quinn & Holland, 1987), retelling is reinterpretation rather than reproduction (Hutcheon, 2006). Advanced video generators can render narrative scenes from text prompts (Ho et al., 2022; Kondratyuk et al., 2024), but rendering a scene in another culture's visual style is not the same as retelling it. We study the latter as **Narrative Cultural Re-instantiation**. Given a source scene and a target cultural context—a documented setting of material forms, social roles, and practices rather than an artistic style—the task is to generate a video in that context while preserving the scene's specified narrative functions and relations (Figure 1).

Current approaches do not directly target this problem. Style-conditioned generators (Liu et al., 2024a; Ye et al., 2025) change how content looks rather than what it is. Commercial video generators prompted with a target culture and culture-aware prompt-refinement methods (Li et al., 2026; Hamna et al., 2025) can change content, but rely on parametric knowledge that is difficult to verify for specific historical contexts. They may also retain source-cultural elements after adaptation. Existing cultural benchmarks instead assess cultural recognizability (Kannen et al., 2024; Nayak et al., 2025; Han et al., 2026), not preservation of a specified source narrative.

Even with reliable cultural knowledge, a second difficulty remains: narrative functions are relational, so substitutes cannot be selected independently. In the poisoned-apple episode of *Snow White*, for example, the object must support offering, acceptance, and consumption; the disguised queen must be replaced by someone who could plausibly approach the heroine; and the setting must make their encounter credible. A substitute may be culturally appropriate in isolation yet fail in combination, such as an iconic object that cannot be consumed or a character whose social role makes the required encounter implausible. Re-instantiation is therefore a *joint structural correspondence* problem rather than a set of independent substitutions.

We propose **CultureWeaver**, a training-free framework that addresses both difficulties by deciding *what* to depict before, and separately from, *how* to render it. To ground cultural knowledge, we construct an evidence-grounded *Visual Cultural Memory* $\mathcal{M}_c = (\mathcal{V}_c, \mathcal{E}_c, \mathcal{I}_c)$, which organizes cultural records $\mathcal{V}_c$, their supported relations $\mathcal{E}_c$, and associated visual evidence $\mathcal{I}_c$. To enforce joint coherence, CultureWeaver abstracts the source scene into a narrative graph, retrieves a candidate graph from $\mathcal{M}_c$, and matches the two using *relation-aware semi-relaxed structural transport*, balancing individual functional suitability with relational and topological compatibility. The resulting cultural bindings and visual references are then realized by frozen image and video generators.

We evaluate CultureWeaver on 200 fairy-tale scenes in Chinese and Persianate target contexts. Task-specific metrics, human evaluation, and ablations show consistent gains in narrative preservation, cultural fidelity, functional validity, and relational coherence. Our contributions are:

- We formulate **Narrative Cultural Re-instantiation** as scene-level cultural reinterpretation under explicit narrative-function and relation constraints.
- We propose **CultureWeaver**, a training-free framework that combines evidence-grounded Visual Cultural Memory with relation-aware semi-relaxed structural transport to determine joint cultural correspondence before rendering.
- We construct a 200-scene, two-context benchmark with culture-neutral annotations and evaluate it through automated, human, and controlled assignment-level analyses.

2 Related Work

Cultural Adaptation and Evaluation. Prior work studies culturally grounded visual generation through transcreation, prompt contextualization, retrieval, and model adaptation (Khanuja et al., 2024; Hamna et al., 2025; Li et al., 2026; Ventura et al., 2025; Liu et al., 2024b). Image transcreation adapts existing content while preserving meaning, whereas KAHANI and MAVEN contextualize characters, scenes, and video prompts for culturally grounded generation. CultureWeaver instead considers source-conditioned re-instantiation, where cultural realizations may change but specified narrative functions and relations must be preserved jointly. Related benchmarks evaluate cultural competence, expectations, inclusivity, and culture-specific content in images and videos (Kannen et al., 2024; Bayramli et al., 2025; Vasilev et al., 2025; Nayak et al., 2025; Han et al., 2026). Our evaluation additionally conditions on a specified source scene, allowing cultural fidelity and narrative preservation to be assessed separately and jointly.

Narrative Generation and Structured Visual Memory. Narrative and reference-guided generation methods improve consistency and control through visual memory, structured layouts, and external visual references (Rahman et al., 2023; Lian et al., 2024; Zhou et al., 2024; Liu et al., 2024a; Ye et al., 2025). Structured representations and retrieval further organize objects, relations, and contextual evidence for generation (Johnson et al., 2018; Khanuja et al., 2024). Visual Cultural Memory builds on these mechanisms to organize alternative cultural realizations together with functional attributes, supported relations, provenance, and visual evidence. Its role is to support cultural correspondence before generation, rather than primarily continuity or rendering control.

Structure-Preserving Correspondence. Gromov–Wasserstein optimal transport compares structured domains through their internal relations (Peyré et al., 2016), while Fused Gromov–Wasserstein (FGW) combines feature and structural information (Vayer et al., 2020). Semi-relaxed formulations constrain one marginal while leaving the other free (Vincent-Cuaz et al., 2022), matching our setting where every source narrative element must be realized but many retrieved candidates may remain unused. In Section 5, we instantiate semi-relaxed FGW with functional compatibility and relation-aware structural costs derived from Visual Cultural Memory, enabling joint correspondence over both node suitability and selected cultural relations.

3 Problem Formulation

We formulate **Narrative Cultural Re-instantiation** as the task of re-expressing a narrative scene in a target cultural context while preserving specified narrative requirements. Given a textual scene description S_{s} and a target-context prompt P_{t}, the goal is to generate a video V_{t} that re-instantiates the scene in the target context, $(S_{\mathsf{s}}, P_{\mathsf{t}}) \mapsto V_{\mathsf{t}}$. The source cultural context is not supplied separately; it is implicit in the characters, objects, settings, and practices described in S_{s}. The prompt P_{t} identifies the intended target cultural context C_{t} without prescribing individual cultural substitutions.

We represent the narrative requirements selected for preservation as a directed, attributed graph $N_{\mathsf{s}} = \mathcal{A}(S_{\mathsf{s}}) = (F_{\mathsf{s}}, R_{\mathsf{s}})$, where $\mathcal{A}$ denotes narrative abstraction openration. Nodes in F_{s} represent salient narrative elements, with attributes specifying their roles, functions, and required affordances where applicable. Relations in R_{s} encode the interactional, causal, temporal, and spatial requirements connecting these elements. A successful output must satisfy two coupled criteria. *Narrative preservation* requires the realized elements and actions to fulfill the specified functions and directed relations in N_{s}. *Cultural coherence* requires their realizations and interactions to remain plausible and interpretable within C_{t}. Neither criterion prescribes a unique target realization, and cultural coherence does not exclude narratively intended deception or transgression.

The criteria are coupled because a candidate that fulfills an individual narrative function may fail to support the required interactions with other selected candidates. Re-instantiation therefore involves *joint structural correspondence*: selecting cultural realizations that jointly support the scene's narrative requirements within the target context. Establishing this correspondence requires contextual evidence about both candidate properties and their relationships. The following section introduces the representation CultureWeaver uses to organize this evidence.

Discussion. Narrative Cultural Re-instantiation differs from cultural generation and appearance-level adaptation in its preservation target. Cultural generation does not require correspondence to a specified source narrative, while appearance-oriented adaptation typically preserves visual content, identity, layout, or motion. Our task instead permits characters, objects, and settings to change, while preserving source narrative functions, affordances, and directed relations. Thus, culturally plausible substitutions may still be invalid when they fail to jointly preserve the narrative structure.

4 Visual Cultural Memory

CultureWeaver externalizes target-cultural knowledge as an evidence-grounded *Visual Cultural Memory* (VCM) $\mathcal{M}_c = (\mathcal{V}_c, \mathcal{E}_c, \mathcal{I}_c)$, where $\mathcal{V}_c$ contains cultural records, $\mathcal{E}_c$ contains supported directed relations among them, and $\mathcal{I}_c$ provides associated visual evidence. The memory is used only at inference time: its structured records and relations support cultural correspondence, while its images support visual realization. Figure 2 summarizes its construction.

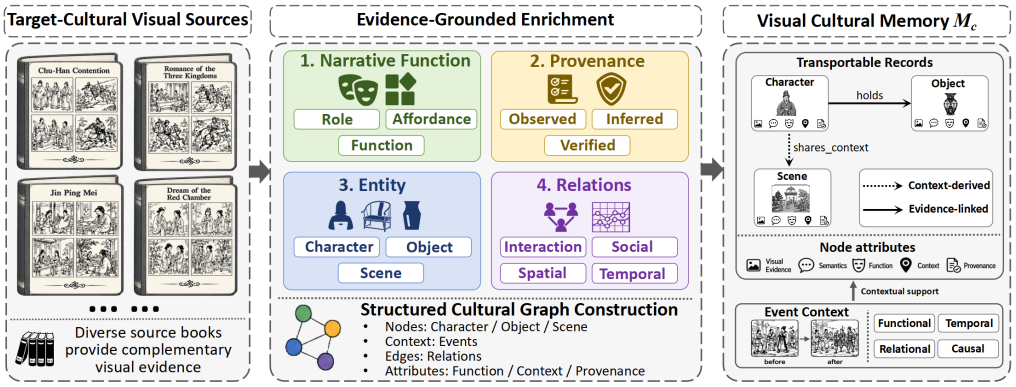

Figure 2: Construction of Visual Cultural Memory. Target-cultural visual sources are enriched with functional, relational, and provenance information and organized into a structured cultural graph.

4.1 Target-Cultural Visual Sources

Each VCM represents a documented target context rather than an undifferentiated cultural label, providing evidence for material forms, social roles, activities, settings, and interactions. We instantiate Chinese and Persianate VCMs as two realizations of the same memory paradigm, using the same construction mechanism over different visual sources. The Chinese VCM contains 848 nodes and 2,412 edges from 552 images across 8 sources, covering represented settings from the late 3rd c. BCE to the Qing period; the Persianate VCM contains 520 nodes and 1,820 edges from 130 images across 4 sources dated ca. 1430–1640 CE. Further details are provided in Appendix J.

4.2 Evidence-Grounded Enrichment

Visual sources and metadata do not fully specify the information required for cultural correspondence. For each source unit, we annotate four complementary components: *narrative function* (role, function, and affordance), *record type* (character, object, or scene), *relations* (interactional, social, spatial, and temporal), and *provenance*. Frozen vision-language and language models propose annotations not directly available from the sources. Each attribute and relation retains its supporting evidence and provenance: Observed is directly supported by visual evidence or source metadata, Inferred is a model-assisted but uncertain interpretation, and Verified is a model-derived annotation accepted through human review. All retained model-derived annotations are manually examined, and unsupported cultural attributions are excluded. Relations are constructed from two evidence sources. *Evidence-linked relations* are directly supported by visual evidence or source annotations, whereas *context-derived relations* are conservatively inferred from local source context under predefined rules. No cross-source relation is introduced without explicit relational evidence.

4.3 Structured Visual Cultural Memory

The resulting record set $\mathcal{V}_c$ contains three transportable types: character, object, and scene. Each record $v_j \in \mathcal{V}_c$ is represented as $v_j = (\tau_j, d_j, z_j, f_j, a_j, x_j, \pi_j, \mathcal{I}_j)$, where τ_j is its record type, d_j its semantic description, z_j its visual attributes, f_j its narrative functions, a_j its affordances, x_j its contextual attributes, π_j its provenance, and $\mathcal{I}_j \subseteq \mathcal{I}_c$ its supporting visual evidence. The relation set $\mathcal{E}_c$ stores supported directed relations among these records, so $(\mathcal{V}_c, \mathcal{E}_c)$ forms a structured cultural graph. A missing edge denotes absence of relational evidence, not cultural incompatibility. Event context is retained at the source-unit level and contributes to contextual attributes, relation interpretation, and retrieval, but is not represented as a record or transport node. Given narrative requirements $N_s = \mathcal{A}(S_s)$, CultureWeaver retrieves records according to required roles, functions, affordances, record types, and incident relations. The retained records and relations form the query-specific candidate graph $G_c^q = (V_c^q, E_c^q)$, with $V_c^q \subseteq \mathcal{V}_c$ and $E_c^q \subseteq \mathcal{E}_c$. Retrieval defines the candidate support but does not determine the final cultural binding; that decision is made by the structural correspondence method in Section 5. Visual evidence associated with the selected records is subsequently used as reference input for visual realization.

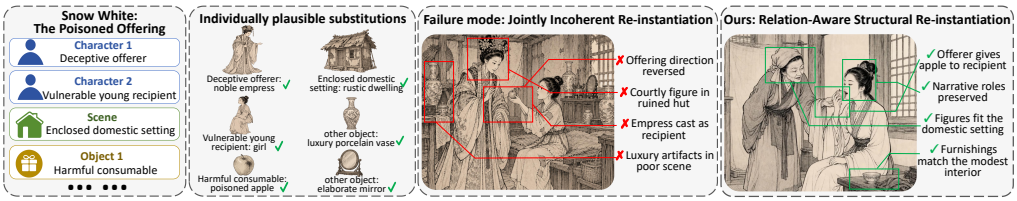

Figure 3: Local plausibility does not imply joint coherence. Individually plausible cultural candidates may become incompatible when combined, motivating joint structural correspondence.

5 Relation-Aware Narrative-to-Culture Structural Transport

Figure 3 illustrates the central challenge: individually plausible cultural candidates may become jointly incompatible when combined. We therefore formulate cultural re-instantiation as correspondence between two structured domains: the narrative graph G_N and the query-specific cultural candidate graph G_c^q. Nodes in G_N specify narrative elements to be culturally realized, whereas nodes in G_c^q provide alternative target-cultural realizations grounded in Visual Cultural Memory. This correspondence is naturally asymmetric: every transportable narrative node requires a realization, while many retrieved candidates may remain unused. We therefore transport importance-weighted mass from narrative nodes to candidate nodes using a semi-relaxed formulation. Node-level costs measure functional suitability, while pairwise structural costs evaluate whether the selected candidates preserve directed relations and graph connectivity. Figure 4 summarizes the framework; implementation details are provided in Appendix I.

5.1 Narrative Requirement Abstraction

We represent the extracted narrative requirements $N_{\mathsf{s}} = \mathcal{A}(S_{\mathsf{s}})$ as a directed attributed graph $G_N = (V_N^{\mathrm{tr}}, E_N)$, produced by a frozen language model under a fixed schema. Nodes in V_N^{tr} correspond to character, object, and scene elements to be realized in the target culture, with attributes specifying their roles, functions, and required affordances. Directed edges in E_N encode relations among these elements. We also retain event context e_N to describe relevant actions and their causal and temporal structure. The abstraction model additionally assigns each transportable node n_i a positive narrative-importance score s_i, used to weight its source-side transport mass.

5.2 Narrative-Conditioned Candidate Retrieval

The target prompt P_{t} selects the corresponding Visual Cultural Memory $\mathcal{M}_{C_{\mathsf{t}}}$. For each narrative node $n_i \in V_N^{\mathrm{tr}}$, CultureWeaver retrieves cultural records using its role, narrative function, affordances, record type, incident relations, and relevant action, causal, and temporal cues from e_N. The retained records and their supported relations form the query-specific candidate graph $G_c^q = (V_c^q, E_c^q)$, where $V_c^q \subseteq \mathcal{V}_{C_{\mathsf{t}}}$ and $E_c^q \subseteq \mathcal{E}_{C_{\mathsf{t}}}$. This stage defines the candidate support for structural correspondence but does not determine the final cultural binding.

5.3 Relation-Aware Structural Correspondence

Given G_N and G_c^q, we seek a soft node correspondence that is compatible both locally and structurally. Local compatibility asks whether a cultural candidate can fulfill the function of a narrative node, while structural compatibility asks whether pairs of assignments preserve the required relations and connectivity. We therefore construct complementary functional and structural costs. For narrative node n_i and cultural candidate c_j, a frozen encoder $\psi(\cdot)$ embeds their functional descriptors, including role, function, and affordance information. The functional cost is

$$M_{ij} = 1 - \cos(\psi(n_i), \psi(c_j)) + \lambda_n D_{\mathrm{node}}(n_i, c_j), \tag{1}$$

where D_{node} captures incompatibilities in role, affordance, record type, and scene context. Node-wise compatibility alone does not ensure a coherent joint assignment. For narrative nodes (n_i, n_r) and an ordered candidate pair (c_j, c_l), we define the structural cost

$$L_{irjl} = \gamma_r \Delta_{\mathrm{rel}}(i, r, j, l) + \gamma_t \left| C_N(i, r) - C_C(j, l) \right|^2, \tag{2}$$

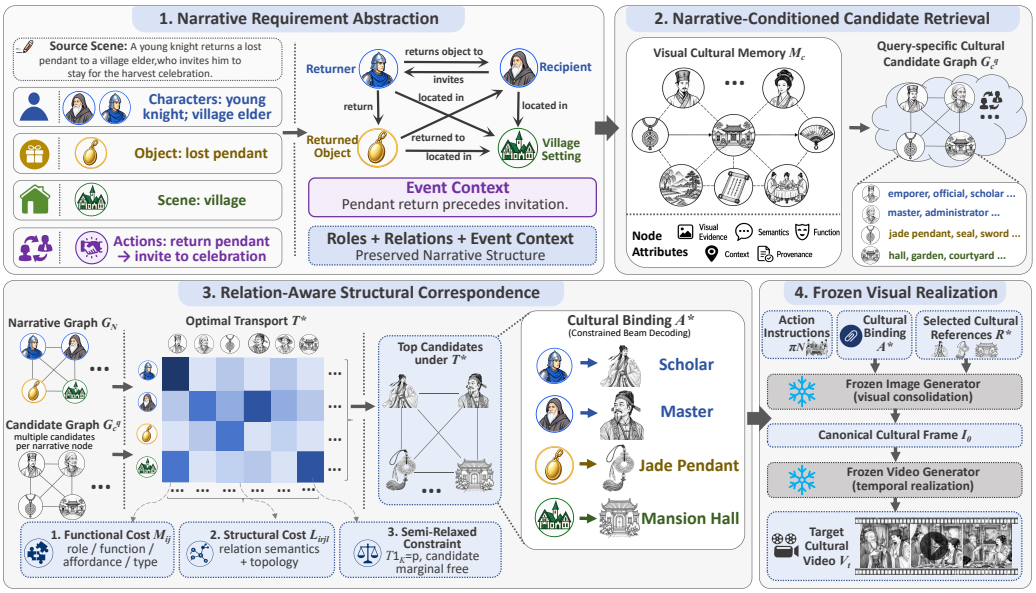

Figure 4: Overview of **CultureWeaver**. Narrative requirements guide candidate retrieval from Visual Cultural Memory, and relation-aware structural transport jointly determines cultural bindings before frozen visual realization.

where Δ_{rel} measures the discrepancy between the required directed relation and the supported candidate relation, including relevant causal, temporal, and evidence-confidence information. C_N and C_C denote normalized directed graph distances that capture relative graph connectivity. When a required target relation lacks supporting evidence, we use a finite penalty rather than treating the candidate pair as strictly incompatible. We instantiate semi-relaxed FGW with these functional and structural costs. Let $m = |V_N^{\text{tr}}|$, $K = |V_c^q|$, and $T \in \mathbb{R}_+^{m \times K}$ denote the soft correspondence, where T_{ij} measures the support for realizing n_i with c_j. We impose $T\mathbf{1}_K = p$, where the importance scores are normalized over transportable nodes as $p_i = s_i / \sum_{n_h \in V_N^{\text{tr}}} s_h$, and leave the candidate-side marginal free. The optimal transport plan is

$$T^* = \arg \min_{\substack{T \geq 0 \\ T\mathbf{1}_K = p}} \left[(1-\alpha) \sum_{i,j} M_{ij} T_{ij} + \alpha \sum_{i,r,j,l} L_{irjl} T_{ij} T_{rl} \right]. \tag{3}$$

The first term favors candidates that satisfy the required narrative functions, while the second couples assignments through relational and topological compatibility. The semi-relaxed constraint preserves the importance-weighted source mass and allows irrelevant candidates to receive negligible mass. We convert T^* into a discrete cultural binding A^* using constrained beam decoding over its highest-mass candidates, enforcing distinct target identities for distinct source-character roles. The decoding score combines transport support and node-level functional cost. Pairwise structural costs are not re-optimized during decoding; their influence is inherited through T^*.

5.4 Frozen Visual Realization

The selected cultural records define the binding A^*, and their associated visual evidence forms the reference set $\mathcal{R}^*$. Narrative action instructions π_N are derived from the preserved relations and event context. A frozen multi-reference image generator combines A^*, $\mathcal{R}^*$, and π_N into a canonical cultural frame $I_0 = F_{\text{img}}(A^*, \mathcal{R}^*, \pi_N)$. A frozen image-to-video generator then produces $V_t = F_{\text{vid}}(I_0, \pi_N)$. Thus, cultural correspondence is determined before generation, while the frozen generators provide visual consolidation and temporal realization.

Table 1: Quantitative comparison on historical Chinese and Persianate cultural re-instantiation using two frozen rendering backends for CultureWeaver. Higher values indicate better performance.

	Historical Chinese						Historical Persianate					
Method	CSD	VQ	NP	CCF	CRV	CRC	CSD	VQ	NP	CCF	CRV	CRC
Seedance 2.0	0.451	4.284	93.1	78.6	80.5	80.7	0.432	**4.315**	91.8	75.3	80.9	80.2
PixVerse V6	0.439	4.127	91.8	76.5	77.9	79.6	0.417	4.136	91.3	74.1	76.8	78.7
Veo 3.1	0.427	4.219	91.5	73.1	78.7	79.8	0.425	4.185	92.1	73.7	78.1	78.7
MAVEN	0.336	4.195	89.7	61.7	74.4	78.3	0.354	4.191	87.1	58.8	73.5	77.0
StyleCrafter	0.413	3.972	84.2	53.3	73.1	73.8	0.377	4.052	83.5	51.5	73.8	72.3
StyleMaster	0.391	4.095	86.6	58.2	70.9	74.2	0.396	4.114	85.8	57.3	71.7	73.1
Ours (Seedance 2.0)	**0.476**	**4.285**	94.9	85.7	86.6	88.1	**0.465**	4.273	94.3	84.1	**86.0**	87.3
Ours (PixVerse V6)	0.468	4.133	**95.7**	**87.3**	**86.7**	**88.9**	0.439	4.158	**94.4**	**85.7**	84.8	**88.1**

6 EXPERIMENTS

Dataset. We evaluate our method on the proposed *Western Fairy-Tale Short-Scene Benchmark*, which contains 200 manually curated short narrative scenes collected from two public-domain European fairy-tale sources: *Household Tales by Brothers Grimm* and *Fairy Tales of Hans Christian Andersen* (Grimm & Grimm, 2004; Andersen, 2008). The benchmark contains 100 Grimm scenes and 100 Andersen scenes, where each source tale contributes at most one scene. Each scene is evaluated under two target cultures: historical Chinese culture and historical Persian culture, resulting in 400 source-target cultural generation conditions. Detailed dataset construction procedures, source-story processing, scene-selection criteria, and benchmark organization are described in Appendix E.

Baselines. We compare against six baselines spanning three categories: commercial video generators with direct cultural prompting (Seedance 2.0, PixVerse V6, and Veo 3.1) (ByteDance Seed, 2026; PixVerse, 2026; Google, 2026b); culture-aware prompt refinement with MAVEN (Li et al., 2026); and reference-guided style-conditioned generation with StyleCrafter and StyleMaster (Liu et al., 2024a; Ye et al., 2025). All methods follow official configurations when available. Commercial models receive the same source scene, target-culture instruction, and culture-specific references. Implementation and fairness details are provided in Appendix F.

Metrics. We evaluate generated videos with CSD for visual style similarity, VQ for technical video quality, and four task-specific metrics: NP for culture-neutral narrative preservation, CCF for target-cultural content fidelity, CRV for functional re-instantiation, and CRC for relational coherence. All four semantic metrics use manually verified, culture-neutral annotations withheld from generation, which specify required functions and relations without fixing a unique cultural realization. CCF, CRV, and CRC additionally use a frozen museum-evidence corpus with generation-memory duplicates removed. Definitions and scoring protocols are provided in Appendix G.

Implementation Details. Unless otherwise specified, GPT-4o is used for language-model-based processing and image generation (OpenAI, 2024). PixVerse V6 is used as the default video-generation backend (PixVerse, 2026). For our method, the cultural assignment module remains unchanged when different video rendering backends are evaluated. All VLM-based metric scores are obtained by averaging Gemini 2.5 Pro (Google, 2026a) and Claude Sonnet 5 (Anthropic, 2026) judgments before aggregation. Additional implementation details are provided in Appendix I.

Main Results. Table 1 shows that CultureWeaver consistently outperforms prior baselines on the core narrative and cultural metrics, including NP, CCF, CRV, and CRC, across both target contexts. The improvements persist with both Seedance 2.0 and PixVerse V6 as frozen rendering backends, indicating that the gains are not tied to a particular video generator. PixVerse V6 achieves the strongest NP, CCF, and CRC results, while Seedance 2.0 yields the highest CSD and Persianate CRV. VQ varies more with the rendering backend, suggesting that CultureWeaver's main advantage lies in narrative preservation and cultural realization rather than low-level video quality. Appendix B reports stable performance across hyperparameter settings. Statistical significance is further assessed through paired-bootstrap comparisons with the strongest baselines for each metric in Appendix C.

Human Evaluation. We conduct blind full-ranking evaluation across all seven methods and convert the rankings into averaged Borda scores. Table 2 shows that CultureWeaver receives the highest preference on narrative and cultural consistency, particularly CCF, CRV, and CRC, while remaining

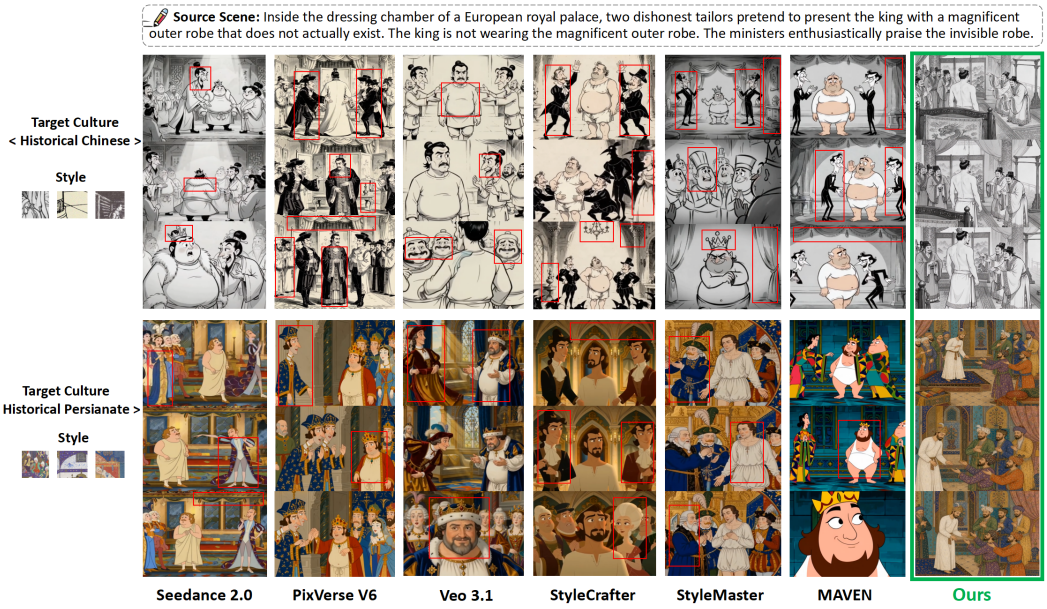

Figure 5: Qualitative comparison with other methods. Our method better preserves the source narrative structure while generating coherent target-cultural characters, objects, and interactions.

Table 2: Blind full-ranking human evaluation results on historical Chinese and historical Persian cultural re-instantiation. Higher Borda scores indicate stronger human preference.

	Historical Chinese						Historical Persianate					
Method	CSD	VQ	NP	CCF	CRV	CRC	CSD	VQ	NP	CCF	CRV	CRC
Seedance 2.0	3.94	**4.97**	4.27	4.48	4.35	4.71	4.15	**4.86**	4.71	4.23	4.54	4.56
PixVerse V6	3.71	4.43	4.31	4.06	3.43	3.59	3.69	4.59	4.88	3.77	4.56	3.42
Veo 3.1	3.64	4.16	3.86	4.21	4.57	4.25	4.02	3.87	3.83	4.05	4.32	4.08
MAVEN	3.14	3.56	3.42	3.14	4.19	4.23	3.01	3.62	3.18	3.28	3.05	3.81
StyleCrafter	4.29	3.32	3.68	3.01	2.71	2.89	3.96	3.40	3.11	3.24	2.53	3.26
StyleMaster	4.57	3.25	3.30	3.03	2.95	2.84	4.24	3.38	3.24	3.16	3.12	3.53
Ours	**4.72**	4.31	**5.17**	**6.06**	**5.83**	**5.46**	**4.81**	4.27	**5.03**	**6.28**	**5.74**	**5.37**

competitive in style similarity and video quality. Table 3 further shows that culture-oriented metrics align more strongly with human judgments than appearance-oriented metrics, with CCF, CRV, and CRC achieving the strongest overall ranking agreement. Together, these results indicate that human judgments and the proposed VLM-based evaluation capture functional and relational aspects beyond visual style adaptation. Human-evaluation details appear in Appendix H.

Qualitative Results. Figure 5 presents qualitative comparisons across different narratives and target cultures. Existing approaches often preserve source-cultural entities or perform only visual stylization, whereas our method generates target-cultural realizations while maintaining the original narrative interactions. Additional qualitative comparisons are provided in Appendix A.

Memory Representation Ablation. Table 4 shows that parametric knowledge alone provides limited cultural grounding, while a flat image–text memory yields only modest improvements. The full Visual Cultural Memory performs best on the narrative and cultural metrics, supporting the value of both external cultural evidence and its structured organization.

Structural Transport Ablation. The lower block isolates the structural correspondence mechanism. Independent Matching performs worst on both pre-render Binding Cost and relation-sensitive video metrics, while removing relation semantics, topology, or semi-relaxation degrades both the intermediate binding objective and downstream CRV/CRC to varying degrees. Full RA-srFGW achieves the strongest overall performance, with the lowest Binding Cost and the best narrative and relational video metrics. Together, these results show that relation-aware structural transport im-

Table 3: Alignment between automatic metrics and human rankings. ρ and τ denote Spearman and Kendall correlations; nDCG@3 and κ_w measure ranking agreement.

Metric	ρ	τ	nDCG@3	κ_w
CSD	0.571	0.476	0.810	0.571
VQ	0.625	0.381	0.762	0.625
NP	0.839	0.667	0.905	0.839
CCF	0.946	0.857	0.956	0.946
CRV	0.839	0.667	0.949	0.839
CRC	0.888	0.795	1.000	0.901

Table 4: Ablation of cultural grounding and structural alignment over two target cultures. Lower Binding Cost denotes better pre-render correspondence under the full RA-srFGW objective.

Method	CSD ↑	VQ ↑	NP ↑	CCF ↑	CRV ↑	CRC ↑	Binding Cost ↓
(A) Cultural Grounding and Memory Representation							
Parametric-only	0.428	4.132	91.6	75.3	77.4	79.2	–
Flat Visual Memory	0.442	4.092	90.5	78.2	79.6	80.4	–
Visual Cultural Memory	**0.454**	**4.146**	**95.1**	**86.5**	**85.8**	**88.5**	–
(B) Structural Alignment and Transport Mechanism							
Independent Matching	0.451	4.126	92.2	83.4	77.8	77.3	0.4254
w/o Relation Semantics	0.446	4.141	92.5	83.7	82.6	78.4	0.4169
w/o Topology	0.449	4.117	93.6	85.1	85.2	86.7	0.4114
w/o Semi-relaxation	0.453	4.124	93.9	85.3	84.5	84.8	0.4102
Full RA-srFGW	**0.454**	**4.146**	**95.1**	**86.5**	**85.8**	**88.5**	**0.4012**

proves correspondence before rendering and that this advantage carries through to the final videos. Binding Cost is an objective-based diagnostic, not an independent cultural correctness metric.

Table 5: RA-srFGW vs. GPT-4o joint assignment under matched inputs and rendering settings. Results are macro-averaged over both target cultures.

Assignment Method	CSD	VQ	NP	CCF	CRV	CRC
GPT-4o Joint Assignment	0.458	4.121	93.3	84.7	84.3	85.8
RA-srFGW	0.454	4.146	95.1	86.5	85.8	88.5

RA-srFGW vs. LLM Joint Assignment. To test whether RA-srFGW can be replaced by general-purpose LLM reasoning, we construct a GPT-4o Joint Assignment baseline with matched narrative graphs, candidate graphs, relational evidence, and rendering settings. As shown in Table 5, both methods achieve similar CSD and VQ, while RA-srFGW consistently improves NP, CCF, CRV, and especially CRC. This indicates that explicit structural optimization provides more coherent functional and relational assignments than direct LLM reasoning under the same cultural evidence. GPT-4o implementation details are provided in Appendix K.

7 Conclusion

We introduced **Narrative Cultural Re-instantiation** as a new perspective on cross-cultural generation: cultural adaptation should preserve what a narrative does, not merely how it looks. Across 200 scenes in Chinese and Persianate target contexts, our method consistently improves narrative and cultural criteria across rendering backends; human evaluation and ablations further support cultural grounding and joint structural correspondence. More broadly, this work shifts cross-cultural generation from appearance-level adaptation toward structure-preserving reinterpretation, where culturally distinct realizations remain valid when they preserve the intended narrative functions and relations without culture-specific parameter adaptation.

AI USE STATEMENT

AI tools were used for language editing and proofreading, including improving grammar, clarity, and phrasing. The research questions, methodological design, experimental protocol, evaluation framework, result analysis, and scientific conclusions were conceived and determined by the authors. AI systems were not used to replace author judgment in the design, interpretation, or validation of the research. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve the collection or use of personally identifiable information, private user data, or sensitive personal attributes. The human evaluation only asks participants to compare and judge outputs produced by different methods, and does not require disclosure of personal information or involve sensitive tasks. The study focuses on cross-cultural narrative generation and is intended for research purposes. We acknowledge that culturally grounded generation may still reflect imperfections or biases in the underlying visual sources and generative models, and therefore interpret the results as evaluations of the proposed method rather than authoritative representations of any culture.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce CultureWeaver and its evaluation throughout the paper, appendices, and supplementary material. An anonymous repository link containing the released code, benchmark, and evaluation resources is provided in the abstract. The benchmark construction, source provenance, manual annotation protocol, and dataset organization are documented in Appendix E. Baseline inputs, configurations, and fairness controls are described in Appendix F, while the definitions and implementation protocols for all evaluation metrics are provided in Appendix G. Model versions, retrieval settings, structural-transport hyperparameters, decoding parameters, generation configurations, and random seeds are specified in Appendix I, and robustness to the principal hyperparameters is reported in Appendix B. The supplementary material includes our implementation code and the experimental Western fairy-tale benchmark constructed from the fixed Grimm and Andersen source editions, together with the associated evaluation annotations and supporting data files.

REFERENCES

Hans Christian Andersen. *Fairy Tales of Hans Christian Andersen*. Project Gutenberg, 2008. URL https://www.gutenberg.org/ebooks/27200. Project Gutenberg eBook No. 27200; last updated January 4, 2021.

Anthropic. Claude models overview: Claude sonnet 5. Claude Platform Documentation, 2026. URL https://platform.claude.com/docs/en/about-claude/models/overview. Accessed 2026-08-20.

Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. Diffusion models through a global lens: Are they culturally inclusive? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 31137–31155. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1503. URL https://aclanthology.org/2025.acl-long.1503/.

ByteDance Seed. Seedance 2.0. https://seed.bytedance.com/en/seedance2_0, 2026. Official model page. Accessed: 2026-08-20.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, New York, NY, USA, 2009. Association for Computing Machinery. doi: 10.1145/1571941.1572114.

Jean-Charles de Borda. *Mémoire sur les Élections au Scrutin*. Histoire de l'Académie Royale des Sciences, 1781.

Bradley Efron and Robert J. Tibshirani. An introduction to the bootstrap. *Chapman & Hall/CRC*, 1994.

Google. Gemini 2.5 pro. Google AI for Developers Documentation, 2026a. URL https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro. Accessed 2026-08-20.

Google. Video generation in the gemini api. https://ai.google.dev/gemini-api/docs/video, 2026b. Official developer documentation for Veo 3.1. Accessed: 2026-08-20.

Jacob Grimm and Wilhelm Grimm. *Household Tales by Brothers Grimm*. Project Gutenberg, 2004. URL https://www.gutenberg.org/ebooks/5314. Project Gutenberg eBook No. 5314; last updated March 1, 2022.

Hamna, Deepthi Sudharsan, Agrima Seth, Ritvik Budhiraja, Deepika Khullar, Vyshak Jain, Kalika Bali, Aditya Vashistha, and Sameer Segal. Kahani: Culturally-nuanced visual storytelling tool for non-western cultures. In *Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '25, pp. 379–400, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3715335.3735478. URL https://doi.org/10.1145/3715335.3735478.

Xianjing Han, Yuhan Su, Yang Deng, Dong Ma, Wee Peng Tay, and Bin Zhu. CultureVidBench: Benchmarking cultural understanding in text-to-video generation. *arXiv preprint arXiv:2608.01942*, 2026. URL https://arxiv.org/abs/2608.01942. Preprint.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. Videoscore2: Think before you score in generated video evaluation. *Transactions on Machine Learning Research*, July 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=MpkVh4jH44.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022. doi: 10.52202/068431-0628. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html.

Linda Hutcheon. *A Theory of Adaptation*. Routledge, 2006. doi: 10.4324/9780203957721.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Johnson_Image_Generation_From_CVPR_2018_paper.html.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/18c669b80d1a8f589713b768bc8fe9a4-Abstract-Datasets_and_Benchmarks_Track.html.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10258–10279. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.573. URL https://aclanthology.org/2024.emnlp-main.573/.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and

Lu Jiang. VideoPoet: A large language model for zero-shot video generation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25105–25124. PMLR, 2024. URL https://proceedings.mlr.press/v235/kondratyuk24a.html.

Shuowei Li, Yuming Zhao, Parth Bhalerao, and Oana Ignat. Maven: A multi-agent framework for multicultural text-to-video generation, 2026.

Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. LLM-grounded video diffusion models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=exKHibougU.

Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Ying Shan, and Yujiu Yang. Stylecrafter: Taming artistic video diffusion with reference-augmented adapter learning. *ACM Transactions on Graphics*, 43(6), 2024a. doi: 10.1145/3687975.

Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. SCoFT: Self-contrastive fine-tuning for equitable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10822–10832, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_SCoFT_Self-Contrastive_Fine-Tuning_for_Equitable_Image_Generation_CVPR_2024_paper.html.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20918–20953. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1141. URL https://aclanthology.org/2025.findings-emnlp.1141/.

OpenAI. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024. Accessed: 2026-08-20.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2664–2672. PMLR, 2016. URL https://proceedings.mlr.press/v48/peyre16.html.

PixVerse. V6 – pixverse platform documentation. https://docs.platform.pixverse.ai/v6-2056814m0, 2026. Official model documentation. Accessed: 2026-08-20.

Vladimir Propp. *Morphology of the folk tale*, volume 10. Univ of TX+ ORM, 2010.

Naomi Quinn and Dorothy Holland. Culture and cognition. In Dorothy Holland and Naomi Quinn (eds.), *Cultural Models in Language and Thought*, pp. 3–40. Cambridge University Press, 1987. doi: 10.1017/CBO9780511607660.002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2493–2502, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Rahman_Make-a-Story_Visual_Memory_Conditioned_Consistent_Story_Generation_CVPR_2023_paper.html.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Investigating style similarity in diffusion models. In *Computer Vision – ECCV 2024*, volume 15124 of *Lecture Notes in Computer Science*, pp. 143–160, Cham, 2025. Springer Nature Switzerland. doi: 10.1007/978-3-031-72848-8_9.

The Metropolitan Museum of Art. The metropolitan museum of art collection api. The Metropolitan Museum of Art Open Access Collection API, 2020. URL https://metmuseum.github.io/. Accessed 2026-08-20.

Viacheslav Vasilev, Julia Agafonova, Nikolai Gerasimenko, Alexander Kapitanov, Polina Mikhailova, Evelina Mironova, and Denis Dimitrov. RusCode: Russian cultural code benchmark for text-to-image generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7656–7672. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.425. URL https://aclanthology.org/2025.findings-naacl.425/.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9), 2020. doi: 10.3390/a13090212. URL https://www.mdpi.com/1999-4893/13/9/212.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models. *Transactions of the Association for Computational Linguistics*, 13:142–166, 2025. doi: 10.1162/tacl_a_00732. URL https://aclanthology.org/2025.tacl-1.10/.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed gromov-wasserstein divergence with applications on graphs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RShaMexjc-x.

Arthur Waley. The chinese cinderella story. *Folklore*, 58(1):226–238, 1947.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2630–2640, 2025. doi: 10.1109/CVPR52734.2025.00251. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ye_StyleMaster_Stylize_Your_Video_with_Artistic_Generation_and_Translation_CVPR_2025_paper.html.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c7138635035501eb71b0adf6ddc319d6-Abstract-Conference.html.

A Additional Qualitative Comparison

Figure 6 provides an additional comparison on the Snow White poisoned-apple scene under both target cultures. Several baselines retain source-cultural visual cues or mainly alter appearance, whereas CultureWeaver re-instantiates the characters and environment while preserving the core offering–consumption–collapse narrative structure. This example further illustrates the distinction between visual adaptation and structure-preserving cultural re-instantiation.

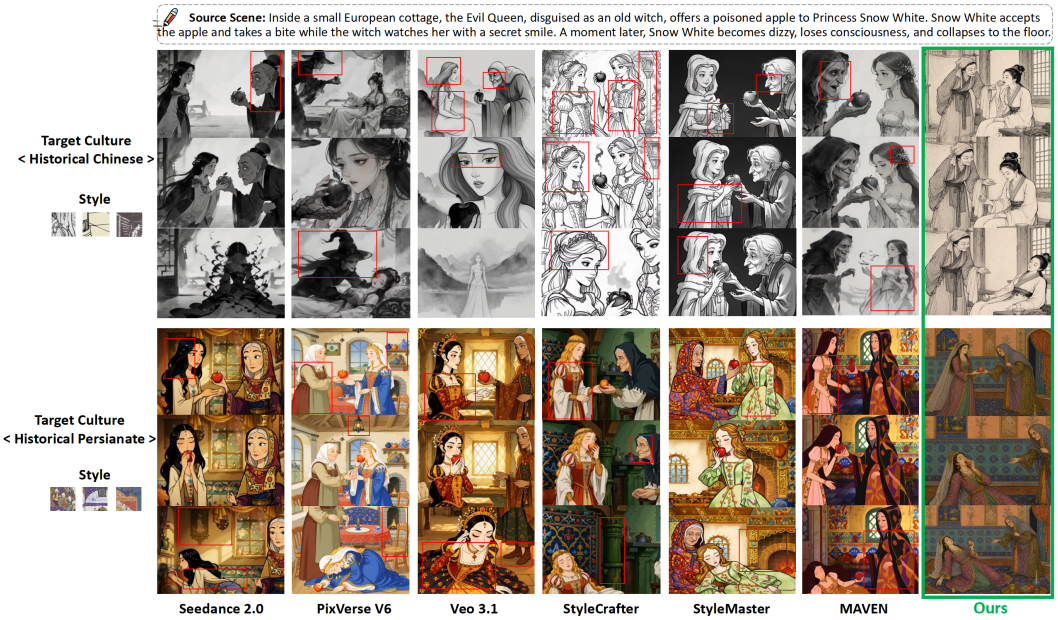

Figure 6: Additional qualitative comparison on the Snow White poisoned-apple scene under historical Chinese and historical Persian cultures. Red boxes mark representative residual or culturally inconsistent visual cues. CultureWeaver better reconstructs target-cultural realizations while preserving the core narrative interaction and outcome.

Table 6: Parameter sensitivity of CultureWeaver. Each parameter is varied independently while all remaining settings are fixed to their defaults. Results are macro-averaged over the two target cultures. Higher is better for all metrics.

Parameter	Setting	CSD	VQ	NP	CCF	CRV	CRC
α	0.2	0.447	4.150	95.3	84.5	81.6	83.3
	0.4	0.450	4.137	95.5	84.8	85.3	85.1
	0.6 (default)	0.454	4.146	95.1	86.5	85.8	88.5
	0.8	0.452	4.142	94.7	84.1	85.1	88.7
γ_{rel}	0.50	0.443	4.145	95.4	85.7	83.1	84.3
	0.65	0.451	4.153	95.2	85.3	83.5	87.4
	0.80 (default)	0.454	4.146	95.1	86.5	85.8	88.5
	0.95	0.448	4.139	95.5	85.8	85.0	88.3
K	6	0.451	4.141	94.4	83.8	84.5	86.9
	12 (default)	0.454	4.146	95.1	86.5	85.8	88.5
	18	0.452	4.157	94.8	86.9	85.3	87.8
	24	0.447	4.136	94.7	85.1	86.4	87.0

B Parameter Sensitivity

Sensitivity Protocol. We evaluate the sensitivity of CultureWeaver to three core hyperparameters: the functional–structural trade-off α, the relation–topology trade-off γ_{rel}, and the number of retrieved candidates per narrative node K. We vary one parameter at a time while keeping the others fixed to their default values. Specifically, we test $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$, $\gamma_{\text{rel}} \in \{0.50, 0.65, 0.80, 0.95\}$ with $\gamma_{\text{top}} = 1 - \gamma_{\text{rel}}$, and $K \in \{6, 12, 18, 24\}$. The default configuration is $\alpha = 0.6$, $\gamma_{\text{rel}} = 0.8$, and $K = 12$.

Analysis. Table 6 shows that CultureWeaver remains stable across a reasonably broad range of hyperparameter settings. For α, the default intermediate weighting provides the best overall balance, while shifting too far toward either functional compatibility or structural consistency leads to less

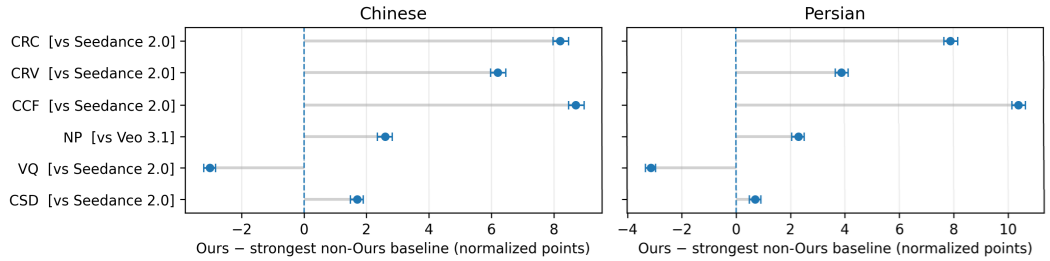

Figure 7: Paired bootstrap comparison with the strongest non-Ours baseline. Points show mean differences with 95% confidence intervals over 10,000 resamples; positive values favor CultureWeaver.

balanced performance. Similarly, $\gamma_{\text{rel}} = 0.8$ achieves the strongest overall trade-off, suggesting that relation semantics and topology provide complementary structural cues. For K, a small candidate set limits cultural coverage, whereas increasing the candidate set beyond the default causes only minor fluctuations, indicating robustness to additional retrieved candidates. Video quality remains largely stable across all settings. Overall, the default configuration provides a robust balance among narrative preservation, cultural grounding, and relational consistency without requiring narrow hyperparameter tuning.

C Paired Bootstrap Analysis

We further compare CultureWeaver with the strongest non-Ours baseline for each metric using paired bootstrap analysis (Efron & Tibshirani, 1994). For each comparison, we perform 10,000 paired resamples and report the mean difference with its 95% confidence interval.

Figure 7 shows positive margins for CSD, NP, CCF, CRV, and CRC across both target cultures, with the largest improvements on CCF, CRV, and CRC. VQ exhibits the opposite trend. This pattern is consistent with the main results: CultureWeaver's gains are concentrated on narrative and cultural criteria rather than low-level video quality.

D Case Study of an Assignment Conflict

This section presents a concrete assignment conflict observed in the Snow White poisoned-apple experiment. All candidates, relations, assignments, and objective values reported below are taken from recorded outputs of the actual matching pipeline rather than manually constructed examples. Independent Matching and RA-srFGW use the same source story, narrative graph, candidate set, functional costs, relation information, and topology matrices. The resulting assignment difference therefore arises from the matching objective rather than from different retrieval results.

Source narrative requirement. The source scene describes a disguised antagonist offering a poisoned apple to a vulnerable young woman in an enclosed encounter setting. Among the relations extracted in the narrative graph is a directed spatial relation from the vulnerable recipient to the encounter setting:

$$v_{\text{recipient}} \xrightarrow{\texttt{located_in}} v_{\text{setting}}. \tag{4}$$

A structurally coherent target-cultural realization should therefore assign the recipient and setting to candidates whose corresponding spatial relation is supported by the target-cultural candidate graph.

Conflict under independent matching. Independent Matching assigns *Receiving Character* to the recipient node and *Palace* to the setting node. Both candidates are individually plausible. *Receiving Character* is described as a character receiving an object with a respectful gesture, while *Palace* is annotated as an architectural setting providing spatial context. For the setting node, *Palace* has a lower node-level functional cost of 0.6395.

However, the two candidates originate from unrelated cultural-memory records, and the candidate graph contains no directed relation from *Receiving Character* to *Palace*. The pair therefore lacks

Table 7: Assignment conflict from the Snow White poisoned-apple experiment. Independent Matching selects the slightly better setting candidate in isolation, whereas RA-srFGW selects a recipient–setting pair supported by the target-cultural candidate graph.

	Independent Matching	RA-srFGW
Recipient	Receiving Character	Receiving Character
Setting	Palace	Pavilion Frame
Setting functional cost	0.6395	0.6551
Same memory record	No	Yes
Required directed relation	Unsupported	`located_in`
Relation confidence	—	0.58

Table 8: Objective decomposition for the complete assignment in the Snow White poisoned-apple conflict case. Component values are reported before weighting; the total objective uses weights 0.40, 0.48, and 0.12 for the functional, relational, and topological components, respectively. Lower values are better.

Component	Weight	Independent	RA-srFGW	Relative change
Functional cost	0.40	0.7484	0.7546	$+0.8\%$
Relation discrepancy	0.48	0.2862	0.2329	-18.6%
Topology discrepancy	0.12	0.1195	0.0918	-23.2%
Total objective	—	0.4511	0.4246	-5.9%

evidence for the required `located_in` relation, despite both candidates being individually reasonable.

Relation-aware correction. RA-srFGW retains *Receiving Character* but changes the setting assignment from *Palace* to *Pavilion Frame*. The latter has a slightly higher node-level functional cost of 0.6551, an absolute increase of only 0.0156. However, *Receiving Character* and *Pavilion Frame* originate from the same cultural-memory record, whose visual evidence depicts the character interacting with others inside a decorated pavilion. More importantly, the target-cultural candidate graph contains the explicit directed relation

$$\textit{Receiving Character} \xrightarrow{\texttt{located_in}} \textit{Pavilion Frame}, \tag{5}$$

with confidence 0.58 and supporting evidence from the associated visual panels. RA-srFGW therefore accepts a small increase in unary functional cost in exchange for a relation-supported and structurally more coherent joint assignment.

Effect on the complete assignment. For comparison, we evaluate both complete discrete assignments under the same RA-srFGW objective decomposition. Independent Matching itself does not optimize this joint objective. Using the default parameters $\alpha = 0.6$, $\gamma_{\text{rel}} = 0.8$, and $\gamma_{\text{top}} = 0.2$, the total objective is

$$\mathcal{J} = 0.40\,\mathcal{L}_{\text{func}} + 0.48\,\mathcal{L}_{\text{rel}} + 0.12\,\mathcal{L}_{\text{top}}. \tag{6}$$

As shown in Table 8, RA-srFGW slightly increases the unweighted functional component from 0.7484 to 0.7546 ($+0.8\%$). In return, the relation discrepancy decreases from 0.2862 to 0.2329 (-18.6%), while the topology discrepancy decreases from 0.1195 to 0.0918 (-23.2%). After applying the objective weights in Eq. 6, the total objective decreases from 0.4511 to 0.4246, corresponding to a 5.9% reduction.

Interpretation. This example illustrates why local plausibility is insufficient for narrative cultural re-instantiation. Independent Matching optimizes the recipient and setting choices separately and therefore favors *Palace* because of its slightly lower unary functional cost, despite the absence of evidence supporting the required recipient–setting relation. RA-srFGW instead evaluates assignments jointly through relation semantics and graph structure. It can therefore sacrifice a small amount of node-level compatibility to obtain a substantially lower relational and topological discrepancy.

The example directly demonstrates the central motivation of CultureWeaver: individually plausible target-cultural substitutions do not necessarily form a coherent realization when combined; their functions and directed relations must also remain jointly compatible.

E WESTERN FAIRY-TALE SHORT-SCENE BENCHMARK

We construct a manually curated benchmark of 200 short narrative scenes derived from two fixed English collections of European fairy tales. One hundred scenes are selected from 100 distinct tales in *Household Tales by Brothers Grimm* (Grimm & Grimm, 2004), and 100 scenes are selected from 100 distinct tales in *Fairy Tales of Hans Christian Andersen* (Andersen, 2008). Each source tale contributes at most one benchmark scene. The final benchmark therefore contains 200 source scenes from 200 distinct tales.

The benchmark is designed for short-form video generation and source-to-target cultural re-instantiation. Each scene represents one temporally compact, visually observable event that can be communicated in approximately 5–10 seconds. The scene descriptions preserve the concrete Western realization of the source narrative while restricting spatial, temporal, and event complexity to a level suitable for short-video generation.

The source-scene benchmark is constructed manually. Story reading, scene selection, event compression, and final inclusion decisions are performed by human annotators rather than by an automatic scene-selection model. Utility scripts are used only for source parsing, data organization, format validation, provenance checking, and descriptive statistics.

In addition to the 200 generation-facing source scenes, the benchmark contains a unified evaluation-only structural annotation for each scene. These annotations support the NP, CCF, CRV, and CRC metrics described in Sec. G. The evaluation annotations are never exposed as input to any video-generation baseline.

E.1 SOURCE COLLECTIONS AND PARSED STORY INVENTORY

We fix the source editions before benchmark construction and do not incorporate modern retellings, online summaries, or alternative translations.

For the Grimm subset, we use Project Gutenberg eBook No. 5314, *Household Tales by Brothers Grimm*, by Jacob Grimm and Wilhelm Grimm, translated by Margaret Hunt (Grimm & Grimm, 2004). The main collection contains tales numbered 1–200, followed by a separate section of Children's Legends. The Children's Legends are excluded from benchmark construction.

Our parser identifies 201 independent Grimm story entries from the main tale section. The additional entry arises from the separately indexed "151* The Twelve Idle Servants", which coexists with the regular numbering sequence. We retain it as an independent candidate source story because it has a distinct title and source-text span.

For the Andersen subset, we use Project Gutenberg eBook No. 27200, *Fairy Tales of Hans Christian Andersen* (Andersen, 2008). The table of contents contains 126 candidate tale entries. We successfully recover 125 corresponding story texts. "The Dumb Cook" appears as a table-of-contents entry but has no corresponding story body in the fixed source file; no missing text is manually reconstructed or imported from another edition.

The parsed source inventory is summarized in Table 9.

Table 9: Source-story inventory used to construct the Western short-scene benchmark. The Grimm count includes the additional indexed tale "151* The Twelve Idle Servants".

Source collection	Candidate entries	Parsed story texts	Final scenes
Brothers Grimm	201	201	100
Hans Christian Andersen	126	125	100
Total	327	326	200

The downloaded source files are preserved without semantic modification. Project Gutenberg header and footer material may be excluded when creating working story-level files, but the original source files remain unchanged.

For each selected benchmark scene, we preserve:

- the Project Gutenberg eBook identifier;
- source collection and story title;
- source-story index or identifier;
- character-level source offsets;
- a contiguous verbatim `source_quote`;
- the SHA-256 checksum of the original source file;
- source URL and retrieval date.

This provenance information makes each benchmark scene traceable to a specific local event in a fixed public source edition.

E.2 Manual Short-Scene Construction

All generation-facing source scenes are constructed manually. Each parsed tale is read in its fixed source edition and inspected for a local narrative event suitable for representation as a short visual sequence. A tale is considered for inclusion only when such an event can be identified without combining distant plot stages.

Scene construction is event-centered rather than summary-centered. The selected scene must correspond to one local event or one tightly connected sequence of actions. Events separated by substantial temporal or spatial distance are not merged into a single benchmark record, and the complete plot of a fairy tale is never compressed into one source scene.

For a selected source tale B_i, we represent the manually constructed scene as

$$c_i = (x_i,\, z_i,\, \mathcal{P}_i,\, \mathcal{O}_i,\, \mathcal{A}_i,\, \ell_i)\,,$$

where x_i is the final generation-facing scene description, z_i is the contiguous source passage supporting the scene, $\mathcal{P}_i$ is the set of salient characters, $\mathcal{O}_i$ is the set of salient objects, $\mathcal{A}_i$ is the ordered set of observable actions, and ℓ_i is the primary physical setting.

Each `source_scene` is written as a self-contained English generation condition. The scenes normally contain 15–45 English words and one to three sentences, with two to four salient characters and/or objects and one to three observable actions.

Literary exposition that is not required to understand the selected event is removed. Manual rewriting is restricted to event compression, clarification of local roles, and conversion into a concise video-generation condition. No unsupported character, object, action, motivation, causal consequence, or target-cultural substitution is added.

The source-domain cultural realization is deliberately retained. Characters, objects, social roles, and environments that carry recognizable Western fairy-tale semantics remain explicit when they are relevant to the selected event. Such realizations include royal roles, courtly or domestic spaces, culturally specific artifacts, magical objects, weapons, gifts, clothing, and social interactions.

For example, one benchmark scene derived from the Snow White narrative is:

> At the seven dwarfs' cottage in a Western kingdom, the Queen, disguised as an old country-woman, offers Snow-white a poisoned red apple. Snow-white takes the poisoned half, bites it, and falls dead while the Queen watches and laughs.

The scene retains source-specific realizations such as the Queen, Snow-white, the dwarfs' cottage, the Western kingdom, and the poisoned apple. No Chinese or Persian target-cultural entity is introduced during dataset construction.

E.3 Manual Scene Quality Control

Candidate scenes are manually evaluated along five dimensions: single-scene coherence, visualizability, interaction strength, cultural re-instantiation potential, and source fidelity. Each dimension is rated on a five-point ordinal scale,

$$q_d(c_i) \in \{1, 2, 3, 4, 5\}.$$

Single-scene coherence. This dimension measures whether the principal event can be represented within one stable or immediately continuous spatial episode. High scores require the salient actions to occur in one local setting without substantial spatial or temporal jumps.

Visualizability. This dimension measures whether the principal narrative semantics can be recovered from visible characters, objects, actions, spatial relations, or observable state changes. Scenes dominated by internal thoughts, exposition, extended dialogue, or unobservable intentions receive lower scores.

Interaction strength. This dimension measures whether the scene contains a meaningful directed interaction or causal relation between salient entities. Representative interactions include offering, giving, commanding, deceiving, testing, threatening, rewarding, rescuing, exchanging, attacking, receiving, hiding, stealing, consuming, or losing an object.

Cultural re-instantiation potential. This dimension measures whether the scene contains roles, objects, settings, actions, or social relations whose concrete cultural realization can change while the underlying narrative function remains stable. Scenes with minimal culture-specific realization receive lower scores.

Source fidelity. This dimension measures whether the normalized scene description is supported by one local source passage. All salient characters, objects, actions, and causal outcomes must remain consistent with the recorded source quotation.

A candidate scene is eligible only when

$$q_{\text{fidelity}} \geq 4, \qquad q_{\text{single}} \geq 4, \qquad q_{\text{visual}} \geq 4,$$

and

$$q_{\text{interaction}} \geq 3, \qquad q_{\text{reinstantiation}} \geq 3.$$

The final description must additionally satisfy the short-video constraints used throughout the benchmark: a compact event sequence, limited action count, and an estimated duration compatible with a 5–10 second generated clip.

All quality judgments used for final dataset inclusion are manually reviewed. Numerical scores serve as structured quality-control records; they do not replace manual scene selection.

E.4 Final Scene Selection and Diversity Control

Final selection is performed independently for the Grimm and Andersen subsets. Exactly 100 scenes are retained from each source collection, with at most one scene selected from any source tale.

Selection considers both scene quality and dataset diversity. The final subset is not obtained by taking the first 100 stories in source order and is not determined by a single scalar ranking criterion.

During construction, we inspect the distribution of scenes across:

- narrative category;
- interaction type;
- physical setting;
- character-role configuration;
- salient object class;

- Western cultural realization;
- concentration of royal or court-centered scenes;
- proportion of dialogue- or exposition-heavy scenes.

The selected benchmark contains narrative events involving, among others, offering and giving, deception, command and hierarchy, testing, threat and punishment, reward, rescue, exchange, conflict, possession and loss, social interaction, ingestion, and visually observable transformation.

Scene settings include royal and domestic interiors, cottages, villages, forests, roads, workshops, rural environments, and supernatural spaces when these are locally supported by the source tale. The character distribution is not restricted to royal protagonists and also contains ordinary household characters, servants, merchants, craftspeople, hunters, children, and anthropomorphic characters where the scene satisfies the benchmark criteria.

When several valid candidates exhibit nearly identical narrative functions, role configurations, or settings, preference is given to the scene that increases coverage of underrepresented configurations.

The final dataset is

$$\mathcal{D} = \mathcal{D}_G \cup \mathcal{D}_A,$$

with

$$|\mathcal{D}_G| = 100, \qquad |\mathcal{D}_A| = 100, \qquad |\mathcal{D}| = 200.$$

E.5 Benchmark Record and Source Traceability

Each selected record contains the generation-facing short scene together with source provenance, structured scene information, and quality-control metadata.

A canonical scene record contains fields equivalent to:

```
{
  "scene_id": "GRIMM_XXX",
  "source_collection": "Household Tales by Brothers Grimm",
  "gutenberg_ebook_id": 5314,
  "source_story_index": 0,
  "story_title": "",

  "source_scene": "",

  "source_span": {
    "start_char": 0,
    "end_char": 0,
    "source_quote": ""
  },

  "characters": [],
  "objects": [],
  "setting": "",

  "actions": [
    {
      "subject": "",
      "predicate": "",
      "object": ""
    }
  ],

  "causal_links": [],
  "temporal_order": [],
  "western_cultural_elements": [],
  "primary_narrative_category": "",
  "estimated_video_duration_sec": 0,

  "quality_scores": {
```

```
    "single_scene": 0,
    "visualizability": 0,
    "interaction_strength": 0,
    "cultural_reinstantiation_potential": 0,
    "source_fidelity": 0
  },

  "accepted": true
}
```


The `source_scene` field is the canonical textual source condition used by downstream video-generation experiments.

The source quotation, source offsets, original story metadata, structured scene fields, and quality-control information are retained for provenance and analysis but are not exposed as additional supervision to baseline video-generation systems.

E.6 Unified Evaluation-Only Structural Annotation

In addition to the generation-facing source scenes, each benchmark record is associated with one unified structural annotation used exclusively for evaluation.

For scene x_i, we denote the annotation as

$$\mathcal{E}_i^{\text{eval}} = \left(\mathcal{V}_i^{\text{eval}}, \mathcal{R}_i^{\text{eval}}, \mathcal{A}_i^{\text{eval}}, \mathcal{U}_i^{\text{eval}}\right),$$

where $\mathcal{V}_i^{\text{eval}}$ contains evaluation nodes, $\mathcal{R}_i^{\text{eval}}$ contains directed relations, $\mathcal{A}_i^{\text{eval}}$ contains narrative atoms, and $\mathcal{U}_i^{\text{eval}}$ contains visual evaluation units.

Evaluation nodes. Each important character, object, or scene entity is represented with:

- `surface_form`;
- entity `type`;
- culture-neutral `narrative_function`;
- required `affordances`;
- `narrative_role`;
- `hard_constraints`;
- `importance`.

These node annotations provide the functional specifications used by CRV.

Relations. Directed relations describe the important interactions among evaluation nodes. Each relation contains:

- subject;
- predicate;
- object;
- direction;
- interaction context;
- causal qualifier;
- temporal qualifier;
- importance.

These relations provide the structural specifications used by CRC.

Narrative atoms. Narrative atoms define the minimal culture-neutral story facts that should remain visually recoverable after cultural re-instantiation. Each atom is assigned an importance weight and is used by NP.

Table 10: Dependence of the six evaluation metrics on the unified evaluation-only annotation.

Metric	Scene-level annotation used
CSD	none
VQ	none
NP	narrative atoms $\mathcal{A}_i^{\text{eval}}$
CCF	visual units $\mathcal{U}_i^{\text{eval}}$
CRV	evaluation-node functions, affordances, constraints, and importance
CRC	directed relations $\mathcal{R}_i^{\text{eval}}$

Visual evaluation units. Visual units identify the characters, objects, or environments whose visible cultural realization is evaluated by CCF. Each unit is associated with a salience weight.

Across the complete 200-scene benchmark, the accepted annotation set contains:

$$801 \text{ evaluation nodes},$$

$$344 \text{ directed relations},$$

$$646 \text{ narrative atoms},$$

and

$$801 \text{ visual evaluation units}.$$

Weights are normalized independently within each scene. For a scene x_i,

$$\sum_{v \in \mathcal{V}_i^{\text{eval}}} p_v = 1, \qquad \sum_{r \in \mathcal{R}_i^{\text{eval}}} w_r = 1,$$

$$\sum_{a \in \mathcal{A}_i^{\text{eval}}} \alpha_a = 1, \qquad \sum_{u \in \mathcal{U}_i^{\text{eval}}} \beta_u = 1.$$

This normalization ensures that scenes containing more annotated entities or relations do not receive greater aggregate weight simply because of annotation density.

The six evaluation metrics depend on these annotations as follows:

E.7 Culture-Neutral Evaluation Annotation

The evaluation annotation specifies the narrative requirements to be preserved but does not prescribe a target-cultural solution.

No evaluation record contains a manually authored Chinese or Persian replacement such as "Chinese palace", "Persian clothing", "Chinese empress", or a specific target-cultural artifact selected as the correct counterpart of a Western source entity.

Instead, the annotation uses culture-neutral functional concepts such as "high-status authority", "harmful consumable", "deceptive offering", "vulnerable recipient", or "royal identity test".

The target culture is introduced only during video generation and during the metric-specific cultural evidence retrieval used by CCF, CRV, and CRC. Consequently, the annotation defines what function or relation should be preserved without defining which Chinese or Persian realization must be used to preserve it.

This separation allows multiple culturally valid target realizations to receive credit and prevents the evaluation protocol from favoring the specific cultural assignment produced by our method.

E.8 Evaluation-Annotation Review and Acceptance

All 200 unified evaluation annotations are individually reviewed before being admitted to the accepted annotation set.

Manual verification covers:

- correctness of node narrative functions and narrative roles;
- validity of affordances and hard constraints;
- direction, causality, and temporal ordering of relations;
- fidelity and completeness of narrative atoms;
- observability and cultural neutrality of visual units;
- consistency of importance and salience assignments.

The annotation workflow preserves intermediate annotation drafts, manual review decisions, and the final accepted records. Only the accepted version is used by the evaluation pipeline.

The benchmark therefore maintains an explicit audit trail separating initial annotation construction from manually verified evaluation specifications.

E.9 Dataset Organization and Core Files

The benchmark is organized under:

```
Dataset/
`-- data/
    `-- western_short_scenes/
        |-- raw/
        |-- parsed/
        |-- candidates/
        |-- selected/
        |   `-- western_short_scenes_200.jsonl
        |-- annotations/
        |   `-- unified_evaluation_annotations_200_accepted.jsonl
        |-- review/
        |   |-- evaluation_annotation_decisions.jsonl
        |   `-- evaluation_annotation_review.csv
        |-- reports/
        |   |-- source_manifest.json
        |   |-- parsing_report.json
        |   |-- validation_report.json
        |   `-- accepted_annotation_validation_report.json
        |-- scripts/
        `-- README.md
```

`raw/`. This directory contains the fixed downloaded Grimm and Andersen source texts and their source-level metadata.

`parsed/`. This directory contains story-level representations obtained by separating the two source collections into independent tales. Parsing is used for source lookup and provenance management and does not determine benchmark membership.

`candidates/`. This directory stores manually constructed candidate scenes prior to final selection, together with source references and scene-level quality-control information.

`selected/`. This directory contains the final 200 generation-facing benchmark scenes. The canonical file is `western_short_scenes_200.jsonl`.

`annotations/`. This directory stores the unified evaluation-only structural annotations. The accepted annotation file used by all VLM-based structural metrics is `unified_evaluation_annotations_200_accepted.jsonl`.

review/. This directory preserves annotation-review records and manual acceptance decisions, including `evaluation_annotation_decisions.jsonl` and `evaluation_annotation_review.csv`.

reports/. This directory stores provenance and validation reports. `source_manifest.json` records source-edition information, `parsing_report.json` records source parsing outcomes, `validation_report.json` records final source-scene validation, and `accepted_annotation_validation_report.json` records validation of the accepted evaluation annotations.

scripts/. This directory contains deterministic utilities for source preparation, story parsing, record validation, identifier checking, descriptive statistics, and report generation. The scripts do not determine the semantic content of the manually constructed source scenes.

README.md. The README documents source editions, dataset fields, construction criteria, directory roles, and validation procedures.

E.10 Validation, Traceability, and Experimental Use

The final benchmark passes validation of:

- final record counts;
- unique scene identifiers;
- source-story uniqueness within each selected subset;
- character offsets into the fixed source files;
- verbatim source quotations;
- scene-length constraints;
- estimated video-duration constraints;
- scene eligibility thresholds;
- completeness of accepted evaluation annotations.

The benchmark is evaluation-only and is not used to train or fine-tune the video generators or the proposed cultural re-instantiation components.

For every source scene x_i, two target-cultural generation conditions are instantiated:

$$(x_i, C_{\text{Chinese}}) \qquad \text{and} \qquad (x_i, C_{\text{Persian}}).$$

The 200 source scenes therefore define

$$200 \times 2 = 400$$

source–target cultural conditions for each evaluated generation method.

For baseline video generation, the canonical `source_scene` is used as the story condition. The complete source tale, supporting source quotation, unified evaluation annotation, and manually reviewed metric specifications are not exposed to the generation model.

The unified evaluation annotation is accessed only after videos have been generated and is used to instantiate NP, CCF, CRV, and CRC. CSD and VQ require no scene-level structural annotation.

Dataset construction, annotation acceptance, and benchmark membership are completed independently of downstream video outputs. Consequently, no scene is added, removed, or rewritten according to the performance of our method or any baseline.

F Baseline Methods and Execution Protocol

We compare our method against six baselines organized into three groups: direct cultural prompting with commercial video-generation models (Seedance 2.0, PixVerse V6, and Veo 3.1), culture-aware prompting (MAVEN), and style-conditioned generation (StyleCrafter and StyleMaster). This

appendix specifies the input construction, operating mode, and fairness constraints used for each baseline.

For each benchmark scene x_i and target culture

$$C \in \{\text{Chinese}, \text{Persian}\},$$

every baseline is evaluated from the same source scene. The only source-story text exposed to all methods is the benchmark field `source_scene`. Story titles, full source tales, supporting quotations, narrative graphs, cultural assignments, retrieved cultural evidence, and evaluation annotations are not exposed to the baselines. Consequently, the comparison measures what each baseline can recover from the same scene-level source condition without access to privileged intermediate structure.

For methods that support image-based style guidance, we define a fixed culture-specific style-reference set

$$R_C = \left\{ r_C^{(1)}, r_C^{(2)} \right\}.$$

For a given target culture, the same two reference images are used for every benchmark scene and every commercial model that supports two-image conditioning. The reference images provide only visual-style and environmental guidance. They contain no story-specific characters, objects, or target assignments.

For methods whose official implementation supports only a single style reference, we define

$$\bar{r}_C = r_C^{(1)}$$

as the canonical style anchor. The same $\bar{r}_C$ is used for all scenes under target culture C.

All baselines are evaluated in a single-pass generation setting. We do not perform scene-specific manual prompt engineering beyond the fixed templates defined below. We do not inspect several semantically different prompt variants and select the most favorable one. When a method exposes random-seed control, the seed is fixed for that baseline and reused across benchmark conditions. When seed control is unavailable, each condition is submitted once and the first successfully returned video is retained. A generation is repeated only in the event of an API failure, corrupted output, or other technical execution failure; semantic dissatisfaction is not a valid reason for regeneration.

For published academic baselines, we follow the authors' released configuration, pretrained models, and recommended inference procedure. For commercial systems, we use their official video-generation interfaces and provide the same source scene, target-culture instruction, and culture-specific style references whenever the corresponding interface supports these inputs. No baseline is given access to the generation-time cultural memory used by our method.

F.1 SEEDANCE 2.0

Seedance 2.0 is used as a direct cultural prompting baseline through the official multimodal video-generation interface provided by ByteDance (ByteDance Seed, 2026). The model supports multimodal conditioning from text and reference media. For the present experiment, we use text and image conditioning only.

For every condition (x_i, C), Seedance 2.0 receives:

- the benchmark source scene x_i;
- the target culture C;
- the two fixed culture-specific style references $r_C^{(1)}$ and $r_C^{(2)}$.

No narrative graph, source-story title, retrieved cultural entity, target-cultural mapping, source video, or evaluation evidence is provided.

The two images are introduced explicitly as style references rather than as literal content references. The model is instructed to preserve the narrative skeleton while replacing the visible source-cultural realization with target-cultural content.

The fixed instruction used for all benchmark scenes is:

```
You are given a short Western source-story scene and two visual style
reference images.

Generate a short video for the supplied scene.

Preserve the original narrative skeleton, including the core action
sequence, directed interactions, and causal outcome.

While preserving this narrative structure, convert the visible content
     of
the scene into <TARGET_CULTURE> cultural realizations. Characters,
clothing, objects, architecture, interior or exterior setting, and
environmental details should be represented in forms appropriate to
<TARGET_CULTURE> culture.

Use the two supplied images as references for the intended visual
    style and
cultural atmosphere.

Do not add unrelated plot events.
Do not expand the scene into the full original fairy tale.
Do not change the central causal relationship of the supplied scene.

Source scene:
<SOURCE_SCENE>
```

Seedance 2.0 therefore measures whether a strong commercial multimodal video model can perform cultural re-instantiation directly from a textual instruction and visual references, without an explicit cultural-memory retrieval stage or structure-preserving assignment mechanism.

F.2 PIXVERSE V6

PixVerse V6 is used as the second direct cultural prompting baseline through the official PixVerse V6 interface (PixVerse, 2026). The V6 platform supports text-to-video, image-to-video, and reference-conditioned video generation. We use the reference-guided generation mode so that PixVerse receives the same type of cultural-style side information as the other commercial multimodal baselines.

For every condition (x_i, C), PixVerse V6 receives:

- the benchmark source scene x_i;
- the target culture C;
- the same two fixed style-reference images R_C.

The source scene and cultural instruction are not rewritten on a per-example basis. No story title, source quotation, narrative graph, retrieved cultural evidence, or manually authored target assignment is provided.

The fixed instruction is:

```
You are given a short Western source-story scene and two visual style
reference images.

Generate a short video while preserving the original story skeleton.

Keep the central action sequence, interaction direction, and causal
    outcome
of the source scene unchanged.
```

```
Convert all visible content into <TARGET_CULTURE> cultural
    realizations,
including characters, attire, props, architecture, spatial environment
    ,
and culturally visible scene details.

Use the two provided images as references for visual style and
    cultural
atmosphere.

Do not introduce additional plot stages.
Do not replace the supplied short scene with a summary of the full
    tale.

Source scene:
<SOURCE_SCENE>
```

PixVerse V6 is evaluated as a direct cultural prompting baseline rather than as a post-hoc video restyling method. Its performance therefore measures whether a proprietary reference-guided video generator can infer an appropriate target-cultural realization directly from the source scene and the cultural instruction.

F.3 VEO 3.1

Veo 3.1 is used as the third direct cultural prompting baseline through Google's official video-generation interface (Google, 2026b). The model supports text-conditioned video generation and image-based direction. We use the image-guided configuration so that the model can receive the same target-culture style references used by the other commercial baselines.

For each condition (x_i, C), the model receives:

- the benchmark source scene x_i;
- the target-culture specification C;
- the fixed reference pair R_C.

The reference images are used only for high-level visual and cultural-style guidance. They do not contain event decomposition, source-to-target entity correspondence, or target-cultural retrieval results.

The fixed instruction is:

```
Generate a short video for the following Western source-story scene.

Preserve the original event structure, interaction direction, and
    causal
outcome.

Transform the visible realization of the scene into
<TARGET_CULTURE> cultural content.
Characters, clothing, objects, architecture, and environmental details
should be represented in culturally appropriate <TARGET_CULTURE> forms
    .

Use the supplied images only as visual-style and cultural-atmosphere
references.

Do not modify the core action sequence.
Do not introduce additional story events.

Source scene:
```

```
<SOURCE_SCENE>
```

Seedance 2.0, PixVerse V6, and Veo 3.1 thus form a controlled direct cultural prompting family. They receive the same benchmark source scene, equivalent cultural instructions, and the same culture-level visual references. No explicit functional abstraction or cultural-entity assignment is performed before generation.

F.4 MAVEN

MAVEN is used as the culture-aware prompting baseline (Li et al., 2026). MAVEN is a multi-agent cultural prompt-refinement framework rather than an independent video foundation model. Its prompt-refinement mechanism decomposes cultural grounding into person, action, and location dimensions. The method studies single-agent refinement, multi-agent sequential refinement, and multi-agent parallel refinement. We use the multi-agent parallel configuration reported as the strongest variant in the original work.

We preserve the generation backend and inference configuration specified by the MAVEN implementation. The refined prompt is rendered using CogVideoX-5B (Yang et al., 2025). Following the reported configuration, videos are generated at 720×480 resolution for 5 seconds at 8 fps, with 50 inference steps, guidance scale 6, and a fixed random seed.

For each condition (x_i, C), the initial textual condition given to MAVEN is:

```
Generate a short video for the following Western source-story scene.

Preserve the core event and causal structure of the supplied scene.

Represent the visible people, actions, objects, and setting using
<TARGET_CULTURE> cultural content.

Scene:
<SOURCE_SCENE>
```

This input is then processed by MAVEN's official multi-agent prompt refinement pipeline. The resulting refined prompt is passed to the original video-generation backend.

MAVEN does not receive the two visual style references used by the commercial multimodal baselines because the published MAVEN framework is a textual cultural-prompt refinement method. Adding external style images would introduce an input modality absent from the original method.

Similarly, MAVEN does not receive:

- the benchmark narrative graph;
- source-node functional signatures;
- source-to-target cultural mappings;
- our cultural-memory retrieval results;
- The Met evaluation evidence.

MAVEN therefore evaluates whether culture-aware prompt decomposition and refinement are sufficient for cultural video generation without an explicit function-preserving and relation-preserving cultural assignment.

F.5 StyleCrafter

StyleCrafter is used as the first style-conditioned video-generation baseline (Liu et al., 2024a). StyleCrafter augments a pretrained text-to-video model with a style-control adapter and explicitly separates textual content conditioning from reference-image style conditioning.

We use the official style-guided text-to-video configuration. No intermediate source video is generated. Consequently, StyleCrafter directly maps a textual source-scene condition and a style image to a stylized video.

Because the published StyleCrafter interface uses a single reference image, we provide the fixed culture-specific canonical style anchor

$$\bar{r}_C = r_C^{(1)}.$$

The same $\bar{r}_C$ is used for all 200 benchmark scenes under culture C.

For every condition (x_i, C), the textual content condition is:

```
Generate a short video for the following Western source-story scene.

Preserve the scene's core action sequence and causal outcome.

Represent the visible characters, clothing, objects, and environment
    using
<TARGET_CULTURE> cultural content.

Use the supplied reference image as the visual-style condition.
Infer the artistic style from the reference image rather than from
additional textual style descriptions.

Scene:
<SOURCE_SCENE>
```

We deliberately avoid adding detailed textual style adjectives because StyleCrafter is designed to obtain style information from its reference image. The textual prompt is therefore primarily responsible for scene content, while the image supplies style conditioning.

No source video, cultural-memory retrieval, narrative graph, or source-to-target cultural assignment is provided.

StyleCrafter therefore measures whether style-guided text-to-video generation can produce culturally transformed content without explicit narrative-level cultural reasoning.

F.6 StyleMaster

StyleMaster is used as the second style-conditioned video baseline (Ye et al., 2025). StyleMaster supports both artistic video generation and video style translation. We use its artistic generation mode rather than its translation mode.

The translation configuration requires a pre-existing source video. Using that configuration would therefore require an additional generator to first synthesize a Western source video, resulting in a two-stage baseline whose performance would depend on both the source-video generator and StyleMaster. To avoid this confound, we use the direct artistic-generation setting.

For each condition (x_i, C), StyleMaster receives:

- the benchmark source scene x_i;
- the target-culture content instruction;
- the canonical culture-specific style image $\bar{r}_C$.

The fixed textual input is:

```
Generate a short video for the following Western source-story scene.

Preserve the original event structure and causal outcome.

Re-render the visible people, clothing, objects, architecture, and
```

```
environment as <TARGET_CULTURE> cultural content.

Use the supplied reference image for visual-style conditioning.

Do not introduce additional plot events.

Scene:
<SOURCE_SCENE>
```

No Western source video is provided, and no additional video stylization stage is applied before or after StyleMaster. The baseline therefore measures direct style-conditioned video generation under the same short-scene setting as the rest of the benchmark.

StyleCrafter and StyleMaster together isolate appearance-level adaptation from narrative-level cultural re-instantiation. They receive explicit visual style information, but they do not explicitly construct a culture-neutral narrative graph, retrieve function-compatible target-cultural entities, or optimize relational compatibility among the selected realizations.

F.7 Fairness and Controlled Information Access

The baseline protocol is designed around controlled information access rather than identical internal inference procedures. Different baselines expose different native input modalities, and forcing all methods into a single artificial interface would alter the methods being compared. We therefore preserve each method's official operating mode while fixing the information content available at the benchmark level.

For every benchmark scene, all methods receive the identical normalized `source_scene`. No method receives the full Grimm or Andersen tale or its title. This restriction reduces the possibility that a model reconstructs additional events from memorized knowledge of a well-known fairy tale.

The three commercial direct-prompting systems receive the same two target-culture style reference images. StyleCrafter and StyleMaster receive the same canonical image selected from that reference pair because their official style-conditioning interface is single-image based. MAVEN receives no style image because its published method operates through textual cultural prompt refinement.

No baseline receives:

- our generation-time visual cultural memory;
- the held-out The Met evaluation corpus;
- retrieved target-cultural candidates;
- our RA-srFGW transport matrix T^*;
- the discrete cultural assignment A^*;
- manually authored source-to-target replacement pairs;
- NP, CCF, CRV, or CRC evaluation annotations.

All methods are evaluated on the same 200 Western short scenes and the same two target cultures. For any baseline b, the complete generation set is therefore

$$\mathcal{V}_b = \{V_b(x_i, C) \mid i = 1, \ldots, 200,\ C \in \{\text{Chinese}, \text{Persian}\}\},$$

containing

$$|\mathcal{V}_b| = 400$$

videos when one successful generation is retained per condition.

With six baselines and our method, the full experimental comparison contains

$$200 \times 2 \times 7 = 2800$$

short videos before any additional ablation experiments.

The same evaluation protocol is applied to all generated videos. In particular, evaluator models are blinded to the generating method, and the evidence retrieved from the held-out The Met corpus for CCF, CRV, and CRC is determined by the source-scene evaluation specification rather than by the method identity or generated output.

The resulting comparison separates three alternative explanations for successful cultural adaptation:

1. direct cultural instruction to a strong multimodal generator;
2. culture-aware textual prompt refinement;
3. reference-guided visual stylization.

Our method differs from these alternatives by explicitly solving for a functionally compatible and relationally coherent target-cultural re-instantiation before visual rendering.

G Evaluation Protocol

We evaluate generated videos along six explicitly separated dimensions: visual style similarity (CSD), technical video quality (VQ), narrative preservation (NP), cultural content fidelity (CCF), cultural re-instantiation validity (CRV), and cultural relational coherence (CRC). The first two metrics are adopted from prior work, whereas the latter four measure complementary requirements of narrative cultural re-instantiation.

Their evaluation responsibilities are deliberately separated. CSD evaluates visual style without judging cultural correctness. VQ measures technical video-generation quality without evaluating narrative or cultural semantics. NP measures whether the culture-independent event structure of the source scene is preserved. CCF measures whether visible characters, objects, and environments are consistent with the specified target culture. CRV measures whether each realized target-cultural element can fulfill the narrative function required by its source role. CRC measures whether the realized elements participate in culturally and socially coherent directed relations.

For every generated video V of duration $T < 10$ seconds, we extract five temporally ordered frames at normalized timestamps

$$\mathcal{F}(V) = \{f_{0.1T}, f_{0.3T}, f_{0.5T}, f_{0.7T}, f_{0.9T}\}.$$

Using interior rather than endpoint frames reduces sensitivity to opening or closing fades while retaining temporal coverage of the entire clip. These five frames are used by CSD and all VLM-based metrics. VideoScore2 is the only exception: we preserve its released full-video inference procedure rather than replacing it with a five-frame approximation.

Unified evaluation-only annotation. Each benchmark scene x_i is associated with one manually verified, culture-neutral evaluation annotation

$$\mathcal{E}_i^{\text{eval}} = \left(\mathcal{V}_i^{\text{eval}}, \mathcal{R}_i^{\text{eval}}, \mathcal{A}_i^{\text{eval}}, \mathcal{U}_i^{\text{eval}}\right),$$

as described in Sec. E.6. Here, $\mathcal{V}_i^{\text{eval}}$ contains evaluation nodes, $\mathcal{R}_i^{\text{eval}}$ contains directed relations, $\mathcal{A}_i^{\text{eval}}$ contains narrative atoms, and $\mathcal{U}_i^{\text{eval}}$ contains visual evaluation units.

The four VLM-based semantic metrics read different parts of this unified annotation:

$$\text{NP} \leftarrow \mathcal{A}_i^{\text{eval}}, \qquad \text{CCF} \leftarrow \mathcal{U}_i^{\text{eval}},$$

$$\text{CRV} \leftarrow \mathcal{V}_i^{\text{eval}}, \qquad \text{CRC} \leftarrow \mathcal{R}_i^{\text{eval}}.$$

The evaluation annotation is completed and manually accepted before any generated video is scored. It contains no Chinese or Persian target assignment and does not specify a unique correct cultural replacement. It is not exposed to Seedance 2.0, PixVerse V6, Veo 3.1, MAVEN, StyleCrafter, StyleMaster, or our generation pipeline.

Table 11: Evaluation dimensions and reported scales. Each metric is assigned a distinct semantic responsibility.

Metric	Evaluated property	Reported scale
CSD	visual style similarity	raw cosine similarity
VQ	technical video quality	VideoScore2 native score
NP	culture-neutral narrative preservation	0–100
CCF	target-cultural visual content fidelity	0–100
CRV	node-level functional re-instantiation	0–100
CRC	relation-level cultural coherence	0–100

VLM judges. All VLM-based judgments are independently produced by Gemini 2.5 Pro (`gemini-2.5-pro`) and Claude Sonnet 5 (`claude-sonnet-5`) (Google, 2026a; Anthropic, 2026). The two judges receive identical textual instructions, frame order, target culture, and external museum evidence.

Method identity, generator-specific prompts, our transport matrix T^*, our cultural assignment A^*, and all generation-time cultural-memory records are removed from evaluator inputs.

For every atomic criterion x, each judge outputs one score from

$$\mathcal{S} = \{0,\ 0.5,\ 1\},$$

where 1 denotes clear visual satisfaction, 0.5 denotes partial or ambiguous satisfaction, and 0 denotes absence, contradiction, or incompatibility.

The two-judge atomic score is

$$\bar{s}_x = \frac{s_x^{\text{Gemini}} + s_x^{\text{Claude}}}{2}.$$

The VLMs never produce the final NP, CCF, CRV, or CRC score. They only produce structured atomic judgments. Final metric values are computed deterministically using the accepted benchmark weights.

For every metric M, results are first averaged over the 200 source scenes separately for each target culture:

$$M_C = \frac{1}{200}\sum_{i=1}^{200} M(V_{i,C}), \qquad C \in \{\text{Chinese}, \text{Persian}\}.$$

We additionally report the culture-macro average

$$M_{\text{Macro}} = \frac{M_{\text{Chinese}} + M_{\text{Persian}}}{2}.$$

G.1 CSD Style Similarity

CSD evaluates whether a generated video follows the intended visual style, independently of whether the depicted entities are narratively or culturally correct. We use the Contrastive Style Descriptor introduced by Somepalli et al. (Somepalli et al., 2025), which was designed to represent image style while reducing dependence on semantic image content.

For each target culture C, we define a fixed two-image style-reference set

$$\mathcal{R}_C = \left\{ r_C^{(1)}, r_C^{(2)} \right\}.$$

The same reference set is used to evaluate every method under culture C. These style references are distinct from the external The Met corpus described in Sec. G.7.

Let $g_{\text{CSD}}(\cdot)$ denote the CSD encoder. We define the normalized descriptor

$$\tilde{g}(x) = \frac{g_{\text{CSD}}(x)}{\|g_{\text{CSD}}(x)\|_2}.$$

The culture-specific style prototype is

$$\mu_C = \frac{\sum_{r \in \mathcal{R}_C} \tilde{g}(r)}{\left\| \sum_{r \in \mathcal{R}_C} \tilde{g}(r) \right\|_2}.$$

For generated video V, the final style score is

$$\mathrm{CSD}(V, C) = \frac{1}{5} \sum_{f \in \mathcal{F}(V)} \tilde{g}(f)^\top \mu_C.$$

We report this mean cosine similarity directly without converting it to a percentage. Higher values indicate stronger agreement with the target visual style.

CSD does not evaluate cultural content. A video that preserves Western characters and artifacts while rendering them in a Chinese or Persian artistic style may therefore receive a high CSD score and simultaneously receive low CCF or CRV scores. This distinction separates visual stylization from cultural re-instantiation.

G.2 Video Quality (VQ)

Technical video quality is measured using the visual-quality dimension of VideoScore2 (He et al., 2026). VideoScore2 separately evaluates visual quality, text-to-video alignment, and physical/common-sense consistency. We retain only its visual-quality output.

Text-to-video alignment is not used because semantic alignment with a source-specific Western description can penalize legitimate cultural replacement and would partially overlap with NP.

We preserve the released VideoScore2 inference procedure. Unlike CSD and the VLM-based metrics, VideoScore2 receives the complete generated video and performs its native temporal sampling.

When a textual condition is required by the released implementation, all methods generated from the same benchmark scene are paired with the same canonical `source_scene`. No method-specific generation prompt or target-cultural assignment is provided to the quality evaluator.

Let ℓ_k denote the model output logit for visual-quality rating $k \in \{1, \ldots, 5\}$. The candidate probabilities are

$$\pi_k = \frac{\exp(\ell_k)}{\sum_{j=1}^{5} \exp(\ell_j)}.$$

Let

$$k^* = \arg \max_{k \in \{1,\ldots,5\}} \pi_k.$$

Following the released soft-score implementation, we report

$$\mathrm{VQ}(V) = k^* \pi_{k^*}.$$

No additional normalization is applied. VQ measures visual clarity, temporal stability, motion quality, and generation artifacts, and is not included in any narrative or cultural metric.

G.3 Narrative Preservation (NP)

Narrative Preservation evaluates whether the generated video preserves the culture-independent event structure of the benchmark source scene. It ignores the literal identity, cultural appearance, costume, architecture, and artistic style used to realize that structure.

For benchmark scene x_i, NP reads the manually verified narrative-atom set

$$\mathcal{A}_i^{\mathrm{eval}} = \{a_{i1}, \ldots, a_{iM_i}\}.$$

Each atom specifies one minimal visually recoverable narrative fact, including directed actions, event occurrence, causal consequences, or temporal dependencies. Culture-specific surface realization is removed from the atom.

For example, a source event involving a queen, a princess, and a poisoned apple may contain atoms such as:

- a powerful antagonist offers a harmful consumable to a vulnerable recipient;
- the recipient accepts or consumes the offered item;
- consumption results in physical harm.

Each narrative atom a_{ij} has a manually verified importance weight α_{ij} satisfying

$$\sum_{j=1}^{M_i} \alpha_{ij} = 1.$$

For atom a_{ij}, judge m returns

$$s_{ij}^{(m)} \in \{0, 0.5, 1\}.$$

The two-judge score is

$$\bar{s}_{ij} = \frac{s_{ij}^{(\text{Gemini})} + s_{ij}^{(\text{Claude})}}{2}.$$

The scene-level NP score is

$$\text{NP}(V_i) = 100 \sum_{j=1}^{M_i} \alpha_{ij} \bar{s}_{ij}.$$

A video can therefore receive a high NP score even if it retains a Western realization, provided that it faithfully preserves the required narrative events. Conversely, visually convincing cultural adaptation receives a low NP score when important events are omitted, reversed, or causally altered.

```
SYSTEM:
You are evaluating Narrative Preservation (NP) for a generated short
video.

You must evaluate only whether the culture-neutral narrative facts
    listed
below are visually supported by the five supplied frames.
The frames are ordered chronologically from FRAME_1 to FRAME_5.

Ignore:
- cultural identity;
- costume;
- architecture;
- artistic style;
- literal character names;
- literal source-object identity.

A culturally different person or object must receive full credit if
    the
same required narrative event is visibly realized.

Do not infer events from prior knowledge of the fairy tale.
Judge only what is supported by the supplied frames.

For each narrative atom, assign exactly one score:

1.0 = the narrative fact is clearly realized;
0.5 = the narrative fact is only partially realized, temporally
    incomplete,
      or visually ambiguous;
0.0 = the narrative fact is absent, contradicted, reversed, or
    incompatible
      with the observed video.

Do not produce an overall NP score.
```

```
Do not evaluate cultural correctness, visual style, or technical video
quality.

INPUT:

Source scene identifier:
<SCENE_ID>

Culture-neutral narrative atoms:
<NARRATIVE_ATOMS_WITH_IDS_AND_IMPORTANCE>

Generated-video frames:
FRAME_1 = 10% duration
FRAME_2 = 30% duration
FRAME_3 = 50% duration
FRAME_4 = 70% duration
FRAME_5 = 90% duration

OUTPUT:

Return valid JSON only:

{
  "atoms": [
    {
      "atom_id": "A1",
      "score": 1.0,
      "evidence_frames": [1, 2],
      "reason": "One concise sentence describing the visible evidence
    ."
    }
  ]
}

Use only 0.0, 0.5, or 1.0 for "score".
```

G.4 CULTURAL CONTENT FIDELITY (CCF)

Cultural Content Fidelity evaluates whether the visible realization of characters, artifacts, and environments is consistent with the specified target culture.

CCF does not evaluate whether a realized entity fulfills the narrative function of its source role; that property is evaluated by CRV. It also does not evaluate whether relations among multiple entities are socially or culturally coherent; that property is evaluated by CRC.

For benchmark scene x_i, CCF reads the manually verified visual-unit set

$$\mathcal{U}_i^{\text{eval}} = \{u_{i1}, \ldots, u_{iL_i}\}.$$

Each visual unit specifies one culturally observable component to be examined in the generated video, including character appearance and attire, salient artifacts, props, environments, or architectural settings.

Visual units are culture-neutral. They describe roles such as "high-status adult female character in an elite interior" rather than target-specific answers such as "Chinese empress" or "Persian noblewoman".

Each unit u_{ij} has a manually verified salience weight β_{ij} satisfying

$$\sum_{j=1}^{L_i} \beta_{ij} = 1.$$

For each visual unit, a method-independent retrieval query is constructed from:

$$q_{\mathrm{CCF}} = [C;\ \text{unit type};\ \text{visual/social role};\ \text{scene context}].$$

The same query and the same retrieved museum evidence are used for all methods evaluated on the same scene, target culture, and visual unit.

The top five records are retrieved from the held-out cultural corpus $\mathcal{M}_C^{\mathrm{eval}}$. Each CCF VLM call therefore receives:

- five chronological frames from the generated video;
- five retrieved The Met images;
- the curator-provided metadata associated with those images.

For visual unit u_{ij}, judge m returns

$$c_{ij}^{(m)} \in \{0, 0.5, 1\},$$

with

$$c_{ij}^{(m)} = \begin{cases} 1, & \text{the unit is clearly visible and culturally consistent,} \\ 0.5, & \text{the unit is visible but culturally ambiguous or partially supported,} \\ 0, & \text{the unit is missing, unidentifiable, or culturally inconsistent.} \end{cases}$$

Missing visual units receive zero because CCF measures realized cultural content rather than cultural correctness conditional on an element being present.

The two-judge unit score is

$$\bar{c}_{ij} = \frac{c_{ij}^{(\mathrm{Gemini})} + c_{ij}^{(\mathrm{Claude})}}{2}.$$

The final scene-level score is

$$\mathrm{CCF}(V_i, C) = 100 \sum_{j=1}^{L_i} \beta_{ij} \bar{c}_{ij}.$$

The retrieved museum examples are treated as non-exhaustive cultural evidence rather than as exact visual targets. A generated entity does not need to duplicate any museum object. The evaluator instead uses the museum image and curator metadata to judge whether the observed visual realization is compatible with the target cultural context.

```
SYSTEM:
You are evaluating Cultural Content Fidelity (CCF) for one predefined
visual unit in a generated short video.

Your task is restricted to one question:

Is the visible realization of this unit culturally consistent with the
specified target culture?

Evaluate only culturally identifiable visual content relevant to the
specified unit, including appearance, attire, artifacts, architecture,
furnishings, and environmental characteristics.

Do NOT evaluate:
- artistic or drawing style;
- technical video quality;
- narrative preservation;
- whether the entity fulfills the source narrative function;
- whether relations among multiple entities are socially coherent.
```

```
The five museum records are independent evidence from The Metropolitan
Museum of Art.
They are non-exhaustive cultural evidence and must not be treated as
    exact
visual templates.

Assign exactly one score:

1.0 = the requested unit is clearly visible and strongly consistent
    with
      the target culture, with no salient contradictory cultural cue;
0.5 = the unit is visible but culturally ambiguous, mixed, partially
      consistent, or weakly supported;
0.0 = the unit is absent, unidentifiable, or visibly inconsistent with
     the
      target culture.

INPUT:

Target culture:
<TARGET_CULTURE>

Evaluation unit:
Unit ID: <UNIT_ID>
Type: <UNIT_TYPE>
Culture-neutral role/context: <UNIT_ROLE_AND_CONTEXT>

Generated-video frames:
FRAME_1 = 10% duration
FRAME_2 = 30% duration
FRAME_3 = 50% duration
FRAME_4 = 70% duration
FRAME_5 = 90% duration

External cultural evidence:
EVIDENCE_1 through EVIDENCE_5

Curatorial metadata:
<MET_EVIDENCE_METADATA_WITH_OBJECT_IDS>

OUTPUT:

Return valid JSON only:

{
  "unit_id": "<UNIT_ID>",
  "score": 0.0,
  "evidence_frames": [],
  "supporting_met_object_ids": [],
  "contradictory_cues": [],
  "reason": "One concise evidence-grounded sentence."
}

Use only 0.0, 0.5, or 1.0 for "score".
```

G.5 Cultural Re-instantiation Validity (CRV)

Cultural Re-instantiation Validity evaluates node-level functional equivalence between a source narrative role and the realization that occupies that role in the generated target-cultural video.

CRV distinguishes genuine functional re-instantiation from both literal preservation of a source-specific entity and culturally plausible but narratively unsuitable substitution.

For scene x_i, CRV reads the accepted evaluation-node set

$$\mathcal{V}_i^{\text{eval}} = \{v_{i1}, \ldots, v_{iN_i}\}.$$

Each node v_{ij} contains a culture-neutral functional signature

$$\phi_{ij} = (t_{ij}, f_{ij}, \mathcal{F}_{ij}, r_{ij}, \mathcal{K}_{ij}),$$

where:

- t_{ij} is the entity type;
- f_{ij} is the narrative function;
- $\mathcal{F}_{ij}$ is the required affordance set;
- r_{ij} is the narrative or social role;
- $\mathcal{K}_{ij}$ is the set of hard functional constraints.

For example, the source-specific "poisoned apple" is evaluated through a functional specification such as an ingestible entity that can be offered, accepted, consumed, conceal harm, and deliver that harm after consumption. The evaluator is not required to find an apple in the target-cultural video.

Each evaluation node has an importance weight p_{ij} satisfying

$$\sum_{j=1}^{N_i} p_{ij} = 1.$$

For each node, a method-independent target-cultural query is generated from

$$q_{\text{CRV}} = [C;\ \text{entity type};\ \text{narrative function};\ \text{required affordances};\ \text{context}].$$

The top five held-out museum records provide external target-cultural context but do not define a unique correct substitute. A realization absent from the retrieved evidence may still receive full credit when its observed function clearly satisfies the accepted functional specification.

For node v_{ij}, judge m outputs

$$r_{ij}^{(m)} \in \{0, 0.5, 1\},$$

where

$$r_{ij}^{(m)} = \begin{cases} 1, & \text{the observed realization satisfies the required function and constraints,} \\ 0.5, & \text{the realization only partially satisfies the functional specification,} \\ 0, & \text{no valid realization is visible or a critical constraint is violated.} \end{cases}$$

CRV does not independently score whether the realization visually belongs to the target culture; that responsibility belongs to CCF.

The two-judge node score is

$$\bar{r}_{ij} = \frac{r_{ij}^{(\text{Gemini})} + r_{ij}^{(\text{Claude})}}{2}.$$

The final score is

$$\text{CRV}(V_i, C) = 100 \sum_{j=1}^{N_i} p_{ij} \bar{r}_{ij}.$$

```
SYSTEM:
You are evaluating Cultural Re-instantiation Validity (CRV) for one
predefined source narrative role.
```

```
Determine whether the entity, object, scene, or event visibly
    occupying
this role in the generated video can perform the culture-neutral
    narrative
function specified below.

Judge functional equivalence rather than literal identity.

Do NOT require the generated realization to reproduce the source-
    specific
entity.
Do NOT reward lexical similarity to the source object.
A different target-cultural realization deserves full credit when it
satisfies the required function, affordances, role, and hard
    constraints.

Do NOT evaluate:
- artistic style;
- technical video quality;
- whether the realization visually looks sufficiently Chinese or
    Persian;
- relation-level coherence among several realized entities.

The five The Met records are external contextual evidence.
They illustrate the target cultural space but do not define the only
    valid
replacement.

Assign exactly one score:

1.0 = a clearly visible realization satisfies the critical narrative
      function, required affordances, and hard constraints;
0.5 = a realization is visible but only partially satisfies the
    functional
      specification, or the required behavior is ambiguous;
0.0 = no valid realization is visible, the functional type is wrong,
    or a
      critical affordance or constraint is violated.

INPUT:

Target culture:
<TARGET_CULTURE>

Evaluation node:
Node ID: <NODE_ID>
Entity type: <ENTITY_TYPE>
Culture-neutral narrative function: <NARRATIVE_FUNCTION>
Required affordances: <AFFORDANCE_LIST>
Narrative/social role: <ROLE>
Hard constraints: <FUNCTIONAL_CONSTRAINTS>

Generated-video frames:
FRAME_1 = 10% duration
FRAME_2 = 30% duration
FRAME_3 = 50% duration
FRAME_4 = 70% duration
FRAME_5 = 90% duration

External target-cultural evidence:
EVIDENCE_1 through EVIDENCE_5

Curatorial metadata:
```

```
<MET_EVIDENCE_METADATA_WITH_OBJECT_IDS>

OUTPUT:

Return valid JSON only:

{
  "node_id": "<NODE_ID>",
  "observed_realization": "A concise factual description.",
  "score": 0.0,
  "satisfied_affordances": [],
  "unsatisfied_affordances": [],
  "evidence_frames": [],
  "supporting_met_object_ids": [],
  "reason": "One concise sentence explaining the functional judgment."
}

Use only 0.0, 0.5, or 1.0 for "score".
```

G.6 CULTURAL RELATIONAL COHERENCE (CRC)

Cultural Relational Coherence evaluates whether the relations instantiated among realized target-cultural elements form a culturally and socially plausible interaction structure.

CRV evaluates individual realizations. CRC instead evaluates whether pairs of realized roles participate in the directed relations required by the source narrative.

For scene x_i, CRC reads the accepted relation set

$$\mathcal{R}_i^{\text{eval}} = \{\rho_{i1}, \ldots, \rho_{iK_i}\}.$$

Each relation ρ_{ij} is represented by the culture-neutral relational specification

$$\psi_{ij} = \left(r_{ij}^s, \delta_{ij}, r_{ij}^o, d_{ij}, s_{ij}, \kappa_{ij}, \tau_{ij}\right),$$

where:

- r_{ij}^s is the subject narrative/social role;
- δ_{ij} is the relation semantics;
- r_{ij}^o is the object or target role;
- d_{ij} is the directed interaction;
- s_{ij} is the interaction or scene context;
- κ_{ij} is the causal qualifier;
- τ_{ij} is the temporal qualifier.

The relation annotation does not contain a target-cultural answer. For example, it may specify that a high-status authority offers an ingestible object to a vulnerable recipient without specifying a Chinese court figure, Persian noble figure, wine cup, or any other target-specific realization.

Each relation has an importance weight ω_{ij} satisfying

$$\sum_{j=1}^{K_i} \omega_{ij} = 1.$$

For each relation, the target-cultural retrieval query is constructed as

$$q_{\text{CRC}} = [C; \text{ subject role; relation semantics; target role; object/action type; scene context}].$$

The retrieval procedure prefers interaction-rich museum records when available, including multi-person paintings, manuscript illustrations, court scenes, domestic scenes, and contextually relevant artifacts.

For relation ρ_{ij}, judge m returns

$$q_{ij}^{(m)} \in \{0, 0.5, 1\},$$

where

$$q_{ij}^{(m)} = \begin{cases} 1, & \text{the directed relation is visible and culturally coherent,} \\ 0.5, & \text{the relation is visible but incomplete, ambiguous, or partially coherent,} \\ 0, & \text{the relation is absent, reversed, or culturally/socially incompatible.} \end{cases}$$

A required relation that is not visibly instantiated receives zero. Relations are never omitted from the denominator according to model output.

The two-judge relation score is

$$\bar{q}_{ij} = \frac{q_{ij}^{(\text{Gemini})} + q_{ij}^{(\text{Claude})}}{2}.$$

The final CRC score is

$$\text{CRC}(V_i, C) = 100 \sum_{j=1}^{K_i} \omega_{ij} \bar{q}_{ij}.$$

```
SYSTEM:
You are evaluating Cultural Relational Coherence (CRC) for one
    predefined
directed relation in a generated target-cultural video.

Determine whether the relation visibly instantiated between the
    specified
roles is culturally and socially coherent in the target culture.

Evaluate only relation-level properties:
- whether the two roles can plausibly participate in this interaction;
- whether the interaction direction is coherent;
- whether the relevant object or action is used appropriately;
- whether social hierarchy is compatible with the interaction;
- whether the scene or context supports the relation;
- whether specified causal or temporal qualifiers are compatible with
    the
  observed sequence.

Do NOT independently evaluate:
- artistic style;
- technical video quality;
- whether each individual entity visually belongs to the target
    culture;
- whether each individual entity separately fulfills its narrative
  function;
- preservation of unrelated parts of the source story.

The five museum records are non-exhaustive external cultural evidence.
They should inform the cultural plausibility judgment but must not be
treated as mandatory exact visual matches.

Assign exactly one score:

1.0 = the intended directed relation is visibly instantiated and
      culturally/socially coherent;
```

```
0.5 = the relation is visible but incomplete, ambiguous, or only
    partly
      coherent in the target context;
0.0 = the relation is absent, reversed, or incompatible with the
    target
      cultural/social context.

INPUT:

Target culture:
<TARGET_CULTURE>

Relation:
Relation ID: <RELATION_ID>
Subject role: <SUBJECT_ROLE>
Relation semantics: <RELATION_SEMANTICS>
Target role: <TARGET_ROLE>
Direction: <DIRECTION>
Scene/context: <RELATIONAL_CONTEXT>
Causal qualifier: <CAUSAL_QUALIFIER>
Temporal qualifier: <TEMPORAL_QUALIFIER>

Generated-video frames:
FRAME_1 = 10% duration
FRAME_2 = 30% duration
FRAME_3 = 50% duration
FRAME_4 = 70% duration
FRAME_5 = 90% duration

External cultural evidence:
EVIDENCE_1 through EVIDENCE_5

Curatorial metadata:
<MET_EVIDENCE_METADATA_WITH_OBJECT_IDS>

OUTPUT:

Return valid JSON only:

{
  "relation_id": "<RELATION_ID>",
  "observed_subject": "Short factual description.",
  "observed_relation": "Short factual description.",
  "observed_object": "Short factual description.",
  "score": 0.0,
  "evidence_frames": [],
  "supporting_met_object_ids": [],
  "coherence_factors": [],
  "conflicting_factors": [],
  "reason": "One concise evidence-grounded sentence."
}

Use only 0.0, 0.5, or 1.0 for "score".
```

G.7 The Met External Cultural Evidence Corpus

CCF, CRV, and CRC use an independent external visual evidence corpus constructed from the Open Access collection of The Metropolitan Museum of Art.

We denote the corpus as

$$\mathcal{M}^{\text{eval}} = \mathcal{M}^{\text{eval}}_{\text{Chinese}} \cup \mathcal{M}^{\text{eval}}_{\text{Persian}},$$

with

$$\left|\mathcal{M}^{\text{eval}}_{\text{Chinese}}\right| = 1000, \qquad \left|\mathcal{M}^{\text{eval}}_{\text{Persian}}\right| = 1000.$$

The resulting evidence corpus therefore contains 2000 museum records. It is constructed and frozen before final video scoring and is never accessible to any generation pipeline, including our cultural-memory retrieval and RA-srFGW modules.

Collection and provenance. Records are collected through The Metropolitan Museum of Art Collection API (The Metropolitan Museum of Art, 2020). Candidate Object IDs are obtained using multiple culture-, geography-, department-, object-, period-, title-, and tag-based queries.

Chinese candidates are retrieved using China/Chinese cultural and geographic terms, with the Asian Art collection serving as a major source. Persian candidates are retrieved using Persian/Iran/Iranian terms, with Islamic Art serving as a major source.

Search membership alone is never treated as a cultural label. Every retained record is validated against curator-provided fields, including:

- `culture`;
- `country`;
- `region`;
- `period`;
- `dynasty`;
- `objectName`;
- `classification`;
- `title`;
- `tags`.

Only records satisfying

$$\texttt{isPublicDomain} = \text{true}$$

and containing a non-empty `primaryImage` are retained.

For every accepted record, we store the original Object ID, primary image, museum-provided metadata, object URL, local image path, and image checksum. Curatorial metadata are preserved verbatim and are never replaced by VLM-generated descriptions.

Evidence diversity. The Chinese and Persian partitions are each constructed to cover multiple types of visual cultural evidence:

$$\begin{aligned}\mathcal{K} = \{&\text{people/attire/social roles},\\ &\text{artifacts/vessels/furniture},\\ &\text{weapons/tools},\\ &\text{architecture/interiors},\\ &\text{court/domestic/public scenes},\\ &\text{ritual/banquet/music/dance},\\ &\text{narrative/multi-person interactions},\\ &\text{landscape/urban context}\}.\end{aligned}$$

Category labels are used for evidence balancing and retrieval support, not as direct evaluation ground truth.

Exact image duplicates are removed using SHA-256 hashes. Near-duplicates are detected through perceptual hashing and manually inspected before one representative is retained.

To prevent evidence leakage, an exclusion manifest containing hashes of generation-memory images is used during corpus construction. Any exact or near-duplicate match is rejected. The generation-time annotations themselves are never exposed to the evaluation corpus builder.

The generation memory and evaluation corpus may therefore cover the same cultural concepts while remaining evidence-disjoint.

Offline retrieval representation. For every museum record e, a text retrieval document is constructed from curator-provided metadata:

$$d_e = [\text{title; objectName; culture; period; dynasty; country; region; classification; tags}].$$

The evidence index does not contain VLM-generated cultural descriptions.

We use hybrid retrieval consisting of lexical BM25 (Robertson & Zaragoza, 2009) and CLIP image-text retrieval (Radford et al., 2021).

For BM25,

$$\mathrm{BM25}(q,d) = \sum_{t \in q} \mathrm{IDF}(t) \frac{f(t,d)(k_1+1)}{f(t,d) + k_1 \left(1 - b + b\frac{|d|}{\overline{L}}\right)},$$

with

$$k_1 = 1.2, \qquad b = 0.75.$$

For visual retrieval, CLIP ViT-L/14 provides text and image representations. For textual query q and museum image I_e,

$$S_{\mathrm{CLIP}}(q,e) = \frac{z_T(q)^\top z_I(I_e)}{\|z_T(q)\|_2 \|z_I(I_e)\|_2}.$$

Because BM25 and CLIP operate on different numerical scales, their raw scores are not directly averaged. Instead, their rankings are fused using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009):

$$\mathrm{RRF}(e \mid q) = \frac{1}{k_0 + r_{\mathrm{BM25}}(e)} + \frac{1}{k_0 + r_{\mathrm{CLIP}}(e)},$$

where

$$k_0 = 60.$$

The top

$$K_{\mathrm{evidence}} = 5$$

records under the fused ranking are supplied to the evaluator.

Metric-specific retrieval. Retrieval queries are constructed from the accepted evaluation annotation, not from the generated video and not from our generated cultural assignment. This ensures that all methods receive identical external evidence for the same scene, target culture, and evaluation unit.

For CCF:

$$q_{\mathrm{CCF}} = [C;\ \text{visual-unit type};\ \text{visual/social role};\ \text{scene context}].$$

For CRV:

$$q_{\mathrm{CRV}} = [C;\ \text{entity type};\ \text{narrative function};\ \text{required affordances};\ \text{context}].$$

For CRC:

$$q_{\mathrm{CRC}} = [C;\ \text{subject role};\ \text{relation semantics};\ \text{target role};\ \text{object/action type};\ \text{scene context}].$$

The queries remain culture-neutral with respect to the target realization. For example, a harmful consumable may be queried through its ingestibility, offering function, and social context rather than through a manually specified Chinese or Persian substitute.

Evidence presentation and evaluator isolation. For every CCF, CRV, or CRC atomic judgment, the evaluator receives ten images in fixed order:

1. the five chronological generated-video frames;
2. the five retrieved The Met evidence images.

Each museum image is accompanied by its Object ID and available curator-provided title, culture, period, dynasty, country, object type, classification, and tags.

The retrieved evidence set is cached before VLM scoring. Gemini and Claude therefore receive identical evidence, and all generation methods receive identical evidence for an identical scene–culture–evaluation-unit tuple.

The evaluator never receives:

- the generation method identity;
- the method-specific generation prompt;
- our transport matrix T^*;
- our cultural assignment A^*;
- generation-time cultural-memory records.

The resulting cultural evaluation is therefore externally grounded while remaining independent of the cultural realization proposed by the generation method being evaluated.

H Human Evaluation Details

We conduct a controlled human evaluation to assess whether the six evaluation dimensions used in our automatic protocol are also reflected in human judgments. The study includes 303 participants from university liberal-arts programs, all of whom are graduate students or faculty members with backgrounds in Chinese cultural studies, Persianate cultural studies, or general humanities and cultural studies.

H.1 Participants and Evaluation Assignment

Among the 303 participants, 113 have backgrounds in Chinese history, literature, and culture, 96 in Persianate cultural studies, and 94 in general humanities and cultural studies. Table 12 summarizes their academic backgrounds and roles.

Table 12: Participant composition by background and academic role.

Background	Faculty	Master's	PhD	Total
Chinese history, literature, and culture	21	55	37	113
Persianate cultural studies	18	53	25	96
General humanities and cultural studies	28	37	29	94
Total	67	145	91	303

Each participant completes eight evaluation questionnaires, each corresponding to one source-scene–target-culture condition. We use a balanced assignment schedule rather than random sampling, ensuring that every condition in the benchmark is evaluated by at least five participants. This yields 2,424 participant-condition evaluations in total, distributed as evenly as possible across the benchmark.

For each condition, participants view outputs from all seven methods generated from the same source scene and target culture. Method identities are hidden and represented by anonymous labels A–G, with the mapping between methods and labels randomized across evaluation pages. For each questionnaire, participants provide six independent full rankings over the seven anonymous videos, one for each evaluation dimension.

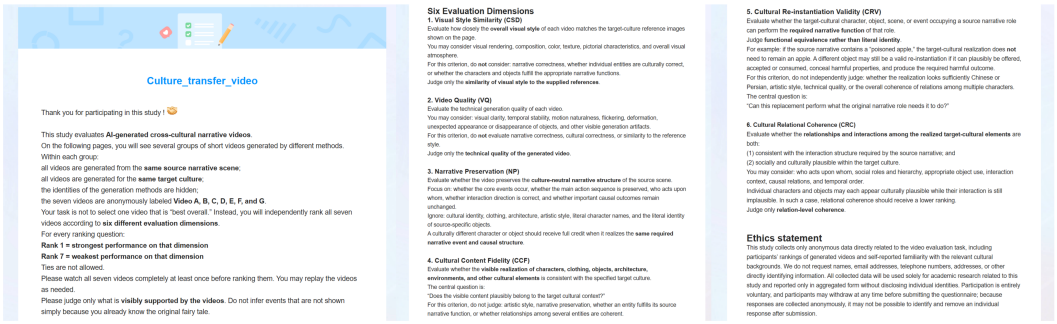

Figure 8: Human-evaluation questionnaire introduction. The page introduces the blind full-ranking task, defines the six evaluation dimensions, and presents the privacy and ethics statement.

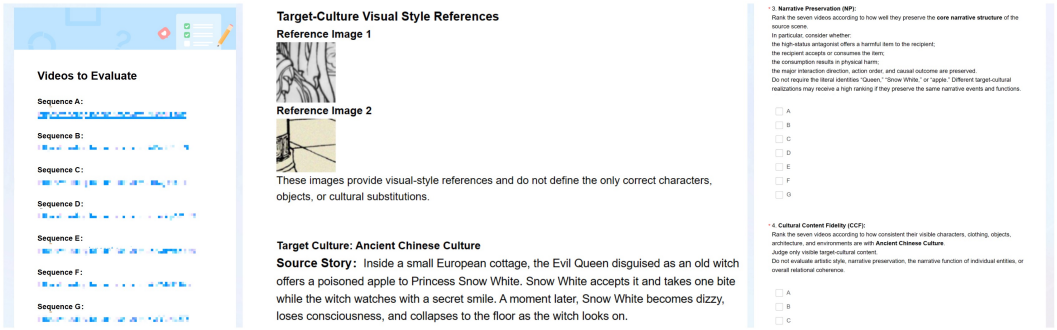

Figure 9: Example human-evaluation page. Participants view seven anonymous videos generated from the same source scene and target culture, together with target-culture style references, and independently rank the videos under each evaluation dimension.

H.2 Questionnaire Design

The questionnaire begins with a task introduction explaining the blind-ranking setting, the six evaluation dimensions, the ranking protocol, and the privacy and ethics statement. Participants are instructed to judge only information visibly supported by the generated videos and not to infer unobserved events from prior knowledge of the original fairy tale. Figure 8 shows the questionnaire introduction page.

Each evaluation page contains: (1) the seven anonymous videos; (2) two target-culture visual-style reference images; (3) the source scene and target culture; and (4) six independent ranking questions. Participants rank all seven videos from best to worst for each dimension, with ties disallowed. Figure 9 shows an example evaluation page.

Evaluation dimensions. The six human-ranking dimensions follow the semantic responsibilities of the automatic metrics, but human rankings are not numerical recomputations of the automatic scores. CSD-aligned human evaluation measures visual-style similarity to the supplied reference images; VQ-aligned evaluation measures technical video quality; NP evaluates preservation of culture-neutral narrative structure; CCF evaluates the visible consistency of characters, clothing, objects, architecture, and environment with the target culture; CRV evaluates whether a target-cultural realization can fulfill the narrative function required by the corresponding source role; and CRC evaluates whether directed interactions among realized elements remain narratively correct and culturally/socially coherent.

In particular, the human CSD and VQ rankings are perceptual judgments rather than outputs of the CSD encoder or VideoScore2. Likewise, human CCF, CRV, and CRC judgments are made directly from the displayed videos and task specification, whereas the corresponding automatic metrics use

structured VLM judgments and external cultural evidence. This separation allows us to test whether the automatic metrics agree with independent human preferences.

Questionnaire template. The questionnaire consists of an instruction page followed by three evaluation pages. The following templates summarize the content presented to participants. Minor layout changes are made by the survey platform, while the wording and evaluation criteria remain fixed.

Part I: Study Introduction
Thank you for participating in this study.
This study evaluates AI-generated cross-cultural narrative videos. On the following pages, you will see several groups of short videos generated by different methods. Within each group:

- all videos are generated from the same source narrative scene;
- all videos are generated for the same target culture;
- the identities of the generation methods are hidden; and
- the seven videos are anonymously labeled Video A–G.

Your task is not to select one video that is "best overall." Instead, you will independently rank all seven videos according to six different evaluation dimensions.
For every ranking question, Rank 1 indicates the strongest performance on that dimension and Rank 7 indicates the weakest performance. Ties are not allowed.
Please watch all seven videos completely at least once before ranking them. You may replay the videos as needed. Judge only what is visibly supported by the videos, and do not infer events that are not shown simply because you already know the original fairy tale.

Six Evaluation Dimensions
1. Visual Style Similarity (CSD). Evaluate how closely the overall visual style of each video matches the target-culture reference images. Consider visual rendering, composition, color, texture, pictorial characteristics, and overall visual atmosphere. Do not judge narrative correctness, cultural-content correctness, or narrative function.
2. Video Quality (VQ). Evaluate the technical generation quality of each video, including visual clarity, temporal stability, motion naturalness, flickering, deformation, and other visible generation artifacts. Do not evaluate narrative or cultural correctness.
3. Narrative Preservation (NP). Evaluate whether the video preserves the culture-neutral narrative structure of the source scene, including the core events, action sequence, directed interactions, and causal outcome. Literal source-character or source-object identity is not required.
4. Cultural Content Fidelity (CCF). Evaluate whether the visible realization of characters, clothing, objects, architecture, environments, and other cultural elements is consistent with the specified target culture. Do not judge artistic style, narrative function, or relation-level coherence.
5. Cultural Re-instantiation Validity (CRV). Evaluate whether the target-cultural character, object, scene, or event occupying a source narrative role can perform the required culture-neutral narrative function of that role. Judge functional equivalence rather than literal identity.
6. Cultural Relational Coherence (CRC). Evaluate whether the relationships and interactions among the realized target-cultural elements both preserve the interaction structure required by the source narrative and remain socially and culturally plausible in the target culture.

Privacy and Ethics Statement
This study collects only anonymous data directly related to the video evaluation task, including participants' rankings of generated videos and self-reported familiarity with the relevant cultural backgrounds. We do not request names, email addresses, telephone numbers, home addresses, or other directly identifying information. All collected data will be used solely for academic research related to this study and reported only in aggregated form without disclosing individual identities. Participation is entirely voluntary, and participants may withdraw at any time before submitting the questionnaire; because responses are collected anonymously, it may not be possible to identify and remove an individual response after submission.

Part II: Evaluation Page
Target culture: `<TARGET_CULTURE>`
Source scene: `<SOURCE_SCENE>`
Two target-culture visual-style reference images are provided. Seven videos generated by different anonymous methods are presented as Video A–G.
For each question below, rank all seven videos from best to worst. Rank 1 indicates the strongest performance on the specified dimension and Rank 7 indicates the weakest. Ties are not allowed.

1. Visual Style Similarity (CSD).
Rank the seven videos according to how closely their overall visual style matches the supplied target-culture reference images. Judge visual style only, rather than narrative correctness or cultural-content correctness.

2. Video Quality (VQ).
Rank the seven videos according to technical generation quality, including visual clarity, temporal stability, motion naturalness, flickering, deformation, and other visible generation artifacts.

3. Narrative Preservation (NP).
Rank the seven videos according to how well they preserve the core culture-neutral narrative structure of the source scene, including the core events, interaction direction, action order, and causal outcome. Different target-cultural characters or objects may receive a high ranking if they preserve the same narrative functions.

4. Cultural Content Fidelity (CCF).
Rank the seven videos according to how consistent their visible characters, clothing, objects, architecture, and environments are with the specified target culture. Judge only visible target-cultural content.

5. Cultural Re-instantiation Validity (CRV).
Rank the seven videos according to whether their target-cultural realizations can perform the narrative functions required by the corresponding source roles. Judge functional equivalence rather than literal identity.

6. Cultural Relational Coherence (CRC).
Rank the seven videos according to whether the relationships and interactions among their realized target-cultural elements preserve the required source interaction structure while remaining socially and culturally plausible in the target culture.

H.3 Borda Aggregation

Each dimension is evaluated using a complete ranking of the seven methods. We convert each ranking into a Borda score (de Borda, 1781). For a method assigned rank $r \in \{1, \dots, 7\}$, its score is

$$b(r) = 8 - r, \tag{7}$$

such that the first-ranked method receives 7 points and the seventh-ranked method receives 1 point.

For method m, evaluation dimension d, and target culture C, the reported human score is the mean Borda score over all human evaluation instances containing culture C:

$$B_{m,d,C} = \frac{1}{|\mathcal{H}_{d,C}|} \sum_{h \in \mathcal{H}_{d,C}} (8 - r_{h,m,d}), \tag{8}$$

where $r_{h,m,d}$ denotes the rank assigned to method m in human evaluation instance h. Higher values indicate stronger human preference.

H.4 Automatic–Human Metric Alignment

We additionally measure whether each automatic metric induces a method ranking consistent with its corresponding human ranking. For each evaluation dimension and target culture, the automatic metric scores and mean human Borda scores independently induce rankings over the seven methods. We compare these two rankings using four complementary statistics.

Spearman correlation (ρ). Spearman correlation measures monotonic agreement between the complete automatic and human method rankings. A value of 1 indicates identical rank order, while lower values indicate increasing disagreement.

Kendall correlation (τ). Kendall's τ measures pairwise ordering consistency. It compares the number of method pairs that are ordered concordantly and discordantly by the automatic and human rankings, and is therefore directly sensitive to pairwise ranking reversals.

nDCG@3. nDCG@3 emphasizes agreement among the highest-ranked methods. The automatic metric determines the predicted method order, while the mean human Borda score is used as the

relevance signal. Let π_i denote the method at automatic rank i and let g_m denote its human Borda score normalized to $[0, 1]$. We compute

$$\mathrm{DCG@3} = \sum_{i=1}^{3} \frac{2^{g_{\pi_i}} - 1}{\log_2(i+1)}, \qquad \mathrm{nDCG@3} = \frac{\mathrm{DCG@3}}{\mathrm{IDCG@3}}, \tag{9}$$

where IDCG@3 is obtained from the ideal ordering induced by human Borda scores.

Weighted Cohen's κ. Weighted Cohen's κ evaluates agreement in the rank position assigned to each method while accounting for agreement expected by chance. We use quadratic weighting, so larger rank disagreements receive larger penalties. For rank categories $i, j \in \{1, \ldots, 7\}$, the disagreement weight is

$$w_{ij} = \frac{(i-j)^2}{(7-1)^2}. \tag{10}$$

This statistic complements correlation measures by directly quantifying agreement in absolute rank assignments.

All four alignment statistics are computed separately for historical Chinese and historical Persian cultural re-instantiation and are then macro-averaged across the two target cultures.

H.5 Privacy and Ethics

The questionnaire collects only anonymous data directly related to the video evaluation task, including participants' rankings and self-reported familiarity with the relevant cultural backgrounds. We do not request names, email addresses, telephone numbers, home addresses, or other directly identifying information. The collected data are used only for academic analysis and are reported in aggregated form without disclosing individual identities. Participation is voluntary, and participants may withdraw before submitting the questionnaire. Because responses are collected anonymously, an individual response may not be identifiable for removal after submission.

I Experimental Settings

Models and implementation. All components are used without parameter updates. We use GPT-4o for narrative graph extraction, frame and video prompt construction, and canonical cultural frame generation. Node-functional and relation-semantic representations are computed with `text-embedding-3-large`. PixVerse V6 is used as the default image-to-video backend in the main experiments, while Seedance 2.0 is additionally evaluated as a supplementary rendering backend. The local pipeline is implemented in Python 3.11.15 on a Linux system equipped with an NVIDIA H100 80GB GPU, using NumPy 2.4.6, SciPy 1.17.1, NetworkX 3.6.1, Pydantic 2.13.4, OpenAI SDK 3.1.0, HTTPX 0.28.1, and Pillow 12.3.0.

Narrative graph construction and candidate retrieval. Each source story is abstracted into a culture-invariant narrative graph whose typed nodes represent character roles, object functions, scene functions, and event functions, and whose directed edges encode interactional, causal, temporal, and spatial relations. For every narrative node, we construct up to four retrieval queries from its narrative function, social role, affordances, and incident relations while excluding source-specific entity names. Full-text search with a BM25-style lexical score is applied over the corresponding functional, role, affordance, and relational descriptions in the cultural memory. We retain the top $K = 12$ candidates per narrative node and construct the candidate graph from their union. Retrieval is used only to provide a high-recall search space; the final cultural correspondence is determined by structural transport.

I.1 Candidate Relations, Missing Evidence, and Topology

Candidate-relation construction. The candidate graph is directed. Explicitly annotated cultural-memory relations are retained whenever both endpoints are included in the retrieved candidate set.

Table 13: Relation evidence used to construct the target-cultural candidate graph. No synthetic relation is introduced across unrelated source units.

Evidence condition	Assigned relation	Confidence
Verified semantic annotation	Annotated relation	0.90
Non-verified semantic annotation	Annotated relation	0.62
Same source unit	`shares_context`	0.35
Same unit, shared discussion action	`discusses_with`	0.58
Same unit, target is spatial setting	`located_in`	0.58
Different units without evidence	No edge	—

For candidates originating from the same visual-cultural source unit, we add conservative context-supported relations according to Table 13. No synthetic relation is introduced between candidates from unrelated source units when no cross-unit relational evidence is available.

The `located_in` relation is directed toward the candidate annotated as an architectural or spatial setting. When multiple supported relations exist for the same ordered candidate pair, the highest-confidence relation is retained in the relation tensor.

Importantly, the absence of an edge does not imply that two cultural candidates are incompatible or cannot coexist. It only indicates that the available cultural memory provides no direct evidence for their relation. Such candidates may therefore still be jointly selected because missing relations receive a finite penalty rather than an infeasibility constraint.

Relation-discrepancy implementation. Let m_{ik}^N and m_{jl}^C indicate whether the directed narrative relation r_{ik}^N and cultural relation r_{jl}^C are present. When both are present, their relation discrepancy is

$$\begin{aligned} D_{\mathrm{rel}}(r_{ik}^N, r_{jl}^C) = 0.62\,\big[1 - \cos\big(\phi(r_{ik}^N), \phi(r_{jl}^C)\big)\big] \\ + 0.10\,\mathbb{I}[a_{ik}^N \neq a_{jl}^C] + 0.10\,\mathbb{I}[t_{ik}^N \neq t_{jl}^C] \\ + 0.08\,\mathbb{I}[d_{ik}^N \neq d_{jl}^C] + 0.10(1 - c_{jl}^C), \end{aligned} \tag{11}$$

where $\phi(\cdot)$ is the frozen relation-description encoder, a and t denote causal and temporal attributes, d denotes stored direction metadata, and c_{jl}^C is the confidence of the cultural relation. The resulting discrepancy is clipped to $[0, 1]$.

Because candidate pairs are ordered, (j, l) and (l, j) correspond to different tensor entries and may therefore receive different relation costs. Practical directionality is thus captured primarily by the ordered source–target indices, together with the stored relation metadata.

The relation discrepancy used by the structural objective is

$$\Delta_{\mathrm{rel}}(i,k,j,l) = \begin{cases} D_{\mathrm{rel}}(r_{ik}^N, r_{jl}^C), & m_{ik}^N = 1,\ m_{jl}^C = 1, \\ 0.90, & m_{ik}^N = 1,\ m_{jl}^C = 0, \\ 0.15(1 - c_{jl}^C), & m_{ik}^N = 0,\ m_{jl}^C = 1, \\ 0, & m_{ik}^N = 0,\ m_{jl}^C = 0. \end{cases} \tag{12}$$

Thus a missing target relation receives a cost of 0.90. This finite penalty represents lack of supporting relational evidence rather than a hard cultural incompatibility. It discourages candidate pairs that fail to support a relation required by the source narrative while still allowing them when their functional compatibility is sufficiently strong or no better-supported alternative is available. Conversely, an additional target-side relation receives only a weaker confidence-dependent penalty.

Directed topology and disconnected candidates. Topology is computed using confidence-weighted directed shortest paths. For a candidate edge e with confidence c_e, its path length is

$$w_e = \frac{1}{\max(c_e, \epsilon)}, \qquad \epsilon = 10^{-6}. \tag{13}$$

Narrative edges are treated analogously using narrative-edge importance. Let $d(a, b)$ denote the directed Dijkstra distance and define

$$d_{\max} = \max\left(1, \max_{d(a,b)<\infty} d(a, b)\right). \tag{14}$$

For an unreachable ordered pair, we assign

$$\widetilde{d}(a, b) = \begin{cases} d(a, b), & d(a, b) < \infty, \\ 1.25\, d_{\max}, & d(a, b) = \infty. \end{cases} \tag{15}$$

The resulting matrix is normalized as

$$C(a, b) = \frac{\widetilde{d}(a, b)}{\max_{x,y} \widetilde{d}(x, y)}. \tag{16}$$

Unreachable ordered pairs therefore receive normalized distance 1, while reachable pairs retain smaller distances according to path length and edge confidence. Disconnected candidates are thus discouraged but are not prohibited from being jointly selected. The topology discrepancy is

$$D_{\text{topo}}(i, k, j, l) = \left(C_N(i, k) - C_C(j, l)\right)^2. \tag{17}$$

Relation and topology costs play complementary roles: the former measures whether a candidate pair realizes the semantics and attributes of a required directed relation, whereas the latter measures whether its surrounding directed connectivity is structurally compatible with the source narrative.

Functional and structural costs. Following Section 5, the node-level functional cost combines cosine distance between frozen semantic representations with an attribute-level discrepancy over narrative function, affordance, social role, entity type, and context. These attribute terms use weights $(0.30, 0.25, 0.20, 0.15, 0.10)$, respectively.

The pairwise structural cost combines the relation discrepancy in Eq. 12 with the topology discrepancy:

$$L_{ikjl} = \gamma_{\text{rel}} \Delta_{\text{rel}}(i, k, j, l) + \gamma_{\text{top}} D_{\text{topo}}(i, k, j, l). \tag{18}$$

Structural transport. We use the relation-aware semi-relaxed Fused Gromov–Wasserstein solver in all main experiments. The complete objective is instantiated as

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{\text{func}} + \alpha\left(\gamma_{\text{rel}}\mathcal{L}_{\text{rel}} + \gamma_{\text{top}}\mathcal{L}_{\text{top}}\right), \tag{19}$$

with $\alpha = 0.6$, $\gamma_{\text{rel}} = 0.8$, and $\gamma_{\text{top}} = 0.2$. The resulting effective weights of the functional, relational, and topological components are therefore 0.40, 0.48, and 0.12, respectively.

Source marginal masses are obtained by normalizing narrative-node importance scores, while the target marginal remains unconstrained, allowing irrelevant candidates to receive negligible transport mass. We optimize the objective with conditional gradients and exact line search for at most 200 iterations, using a convergence tolerance of 10^{-8}. Each example is solved with three restarts: one functionally initialized run and two random initializations. We use seed 0, producing restart seeds $\{0, 1, 2\}$, and retain the solution with the lowest objective value.

Transport decoding and reference selection. After obtaining the optimal transport plan T^*, we retain the five candidates with the largest transport mass for each narrative node. A beam search with width 64 converts the soft correspondence into a discrete cultural binding. The decoding score combines negative transport mass with node-level functional cost, weighted by 0.15. For character nodes, we additionally enforce an identity constraint that prevents the same target-cultural character identity from filling multiple source-character roles.

The decoder does not explicitly re-evaluate pairwise relations among selected candidates; relation and topology information have already shaped the transport plan through RA-srFGW. We then select at most six distinct reference images, ranked by transport support and narrative-node importance, to form the reference set for canonical cultural frame generation.

Image and video generation. Canonical cultural frames are generated with GPT-4o at 1024×1024 resolution using medium quality and the selected cultural reference images. Reference images are resized to a maximum side length of 1536 pixels and encoded as JPEG with quality 88 before upload. Up to three image-generation attempts are allowed, with progressively reduced reference resolution and compression quality following API or reference-upload failures.

PixVerse V6 is used as the default video-generation backend with normal motion mode, 720p resolution, an 8-second duration, and seed 0. For supplementary backend comparison, we additionally use Seedance 2.0 with the same target video settings: 8-second duration, 720p resolution, normal motion configuration, and seed 0. The cultural binding, selected references, and canonical cultural frame are kept fixed when changing the video backend, so the comparison isolates the effect of the final video renderer.

Table 14: Fixed implementation parameters used in the main experiments.

Parameter	Setting
Language/image model	GPT-4o
Node/relation embedding model	`text-embedding-3-large`
Candidates per narrative node K	12
Node-attribute weights	$(0.30, 0.25, 0.20, 0.15, 0.10)$
Relation-component weights	$(0.62, 0.10, 0.10, 0.08, 0.10)$
Missing-relation cost	0.90
Functional–structural weight α	0.60
Relation weight γ_{rel}	0.80
Topology weight γ_{top}	0.20
Maximum transport iterations	200
Convergence tolerance	10^{-8}
Transport restarts	3
Transport initialization	Functional, then random
Base random seed	0
Decoding candidates per node	5
Decoding beam width	64
Functional decoding coefficient	0.15
Maximum reference images	6
Canonical-frame resolution	1024×1024
Canonical-frame quality	Medium
Image-generation attempts	3
Maximum reference-image side	1536 pixels
Reference JPEG quality	88
Default video backend	PixVerse V6
Supplementary video backend	Seedance 2.0
Video duration and quality	8 seconds, 720p
Video motion mode	Normal
Video seed	0

Fixed hyperparameters and reproducibility. Table 14 summarizes the fixed settings used across all stories and both target cultures. We do not tune these values separately for Chinese and Persian experiments; the target culture is changed by substituting the corresponding Visual Cultural Memory while keeping the narrative representation, structural transport, and decoding procedure unchanged. Narrative graphs, retrieved candidates, cost tensors, transport plans, decoded bindings,

Table 15: Source and scale summary of the two generation-time Visual Cultural Memories. Chinese temporal coverage refers to historical and literary settings represented by the source narratives, whereas Persianate temporal coverage refers approximately to production dates of the underlying artworks. Human review denotes the proportion of retained model-derived annotations that were manually examined; it is not the proportion marked VERIFIED and does not estimate annotation accuracy.

Cultural memory	Sources	Units	Images	Transportable records	Source type	Temporal coverage	Human review (%)
Chinese	8	276	552	848	*Lianhuanhua*	late 3rd c. BCE–Qing-period settings	100
Persianate	4	130	130	520	Persianate paintings / manuscript illustrations	ca. 1430–1640 CE	100

selected references, prompts, generated frames, video-task identifiers, and run configurations are stored in per-example run directories. Interrupted API stages can therefore be resumed without recomputing completed stages. Random seeds control the local transport restarts and video-generation requests; hosted GPT-4o inference otherwise follows the provider defaults, and returned artifacts are cached to support reproducibility.

J VISUAL CULTURAL MEMORY DETAILS

This appendix provides source provenance, annotation procedures, inventory statistics, and candidate-graph construction details for the Visual Cultural Memory described in Section 4. We instantiate Chinese and Persianate VCMs as two concrete realizations of the same memory paradigm. Both are external inference-time cultural knowledge resources; their images and annotations are not used to update model parameters.

J.1 INSTANTIATION AND SOURCE PROVENANCE

Chinese VCM. The Chinese VCM is constructed from eight *lianhuanhua* narrative collections and contains 276 paired narrative units comprising 552 images. The represented historical and literary settings span from the Chu–Han transition in the late third century BCE to the Qing-period social world represented in *Dream of the Red Chamber*. This temporal range describes the settings represented by the narratives, rather than the publication dates of the *lianhuanhua* editions.

Persianate VCM. The Persianate VCM contains 130 visual records collected through the Naar-Doon library of Eranshahr. The selected records are traced to four upstream visual sources: Christie's, Islamoriente, The David Collection, and the Los Angeles County Museum of Art (LACMA). The underlying Persianate paintings and manuscript illustrations date approximately from 1430 to 1640 CE, covering Timurid Herat, Safavid-era traditions associated with Tabriz, Shiraz, and Isfahan, and related sixteenth-century Persianate painting traditions. Record-level provenance to the underlying artwork and upstream source is retained.

The two VCMs differ in source format and temporal interpretation. Chinese source units consist of paired narrative panels, and their temporal coverage refers to represented historical and literary settings. Persianate source units correspond to individual visual records, and their temporal coverage refers approximately to artwork production dates. The two memories are constructed independently and contain no manually specified source-to-target correspondences.

J.2 RECORD SCHEMA, EVENT CONTEXT, AND MEMORY INVENTORY

Consistent with Section 4.3, the record set $\mathcal{V}_c$ contains three transportable record types: character, object, and scene. Each source unit additionally retains *event context* describing relevant actions and functional, relational, causal, or temporal structure. Event context is auxiliary contextual information rather than a transport node. It may enrich record attributes, relation interpretation, and retrieval, but does not receive an independent source-to-target cultural binding.

Each transportable record follows the representation

$$v_j = (\tau_j, d_j, z_j, f_j, a_j, x_j, \pi_j, \mathcal{I}_j), \tag{20}$$

Table 16: Detailed inventory of the two generation-time Visual Cultural Memories. Transportable records comprise character, object, and scene records. Event-context annotations are stored separately as auxiliary functional, relational, causal, and temporal information and are not transport nodes. Raw relations denote semantic-relation annotations before transportable-endpoint filtering and triplet-level deduplication.

Cultural memory	Units	Images	Transportable records	Characters	Objects	Scenes	Event-context annotations	Raw relations
Chinese	276	552	848	540	112	196	552	1,144
Persianate	130	130	520	268	112	140	130	650

where τ_j denotes its record type, d_j its semantic description, z_j its visual attributes, f_j its narrative functions, a_j its affordances, x_j its contextual attributes, π_j its provenance, and $\mathcal{I}_j \subseteq \mathcal{I}_c$ its supporting visual evidence. Original captions and source metadata are preserved whenever available.

Normalized record categories. For reporting and inference-time retrieval, source-specific entity labels are normalized into the three shared transportable categories.

In the Chinese VCM, character records include `main_character`, `supporting_character`, `crowd`, and `animal`; object records include `prop`, `weapon`, `vehicle`, and `other`; and scene records include `scene_landmark`, `background_entity`, and `building`.

In the Persianate VCM, character records include `character`, `character_group`, `animal`, `animal_group`, and `creature`; object records correspond to `object`; and scene records include `scene` and `natural_feature`.

Normalization provides a shared correspondence interface while retaining source-specific descriptions, visual attributes, contextual annotations, and supporting evidence.

The source collections contain heterogeneous characters, social roles, artifacts, architecture, domestic and public environments, landscapes, activities, and multi-entity interactions. This diversity provides visual and contextual evidence for a range of narrative functions and relations rather than concentrating the memory on a single scene type or visual category.

J.3 Annotation Enrichment, Provenance, and Human Review

Source images, captions, and metadata do not exhaustively specify narrative functions, affordances, social roles, interaction semantics, or contextual information. Both memories therefore use the same frozen vision-language and language-model enrichment protocol described in Section 4. Model-assisted enrichment proposes semantic descriptions, visual attributes, narrative functions, affordances, contextual information, event context, and semantic relations. For model-derived annotations, supporting image identifiers, confidence information, model identifiers, and annotation versions are retained for traceability.

Provenance status is assigned at the attribute and relation level:

- Observed denotes information directly supported by visible evidence, source annotations, captions, or metadata;
- Inferred denotes information obtained through model-assisted visual or contextual interpretation that remains uncertain or unconfirmed;
- Verified denotes a model-derived annotation accepted through human review against its supporting evidence and local context.

These states describe evidence status rather than a mandatory sequential pipeline. Source-supported information can be recorded directly as Observed, while review of a model-derived annotation may result in Verified, retention as Inferred, or exclusion.

Human review protocol. All model-derived annotations retained in the final generation-time VCMs are manually reviewed. Reviewers inspect each proposed annotation together with its supporting visual evidence, available source metadata, and local narrative context. The review considers whether the proposed record interpretation, cultural attribution, narrative function, affordance,

contextual annotation, or relation semantics is supported by the available evidence. Annotations accepted by reviewers are marked VERIFIED; ambiguous or unconfirmed annotations remain INFERRED; and unsupported cultural attributions are excluded from the memory.

We define review coverage as

$$R_c^{\text{review}} = \frac{N_{\text{reviewed}}^c}{N_{\text{model-derived}}^c} \times 100\%, \tag{21}$$

where $N_{\text{model-derived}}^c$ counts retained model-derived annotations and N_{reviewed}^c counts those subjected to human review. For both generation-time memories, $R_c^{\text{review}} = 100\%$. This quantity measures review coverage only; it is neither the proportion of annotations marked VERIFIED nor an estimate of annotation accuracy.

Cultural content and visual style. We separately retain evidence concerning represented cultural content and visual presentation. Annotations of clothing, artifacts, roles, activities, settings, and interactions concern represented content, whereas palette, line work, brushwork, and illustration technique concern visual style. This distinction prevents stylistic resemblance alone from being treated as evidence that a record satisfies the required cultural or narrative function.

J.4 CANDIDATE-GRAPH MATERIALIZATION

The complete VCM and the query-specific candidate graph are different representations of the same cultural knowledge. The complete memory retains the full set of transportable records, supported relations, contextual annotations, provenance, and visual evidence. At inference time, records are retrieved according to narrative functions, affordances, role constraints, record types, and incident relations.

For a query q, the retained records and their supported relations form

$$G_c^q = (V_c^q, E_c^q), \qquad V_c^q \subseteq \mathcal{V}_c, \qquad E_c^q \subseteq \mathcal{E}_c. \tag{22}$$

More specifically,

$$E_c^q = \{(u, r, v) \in \mathcal{E}_c \mid u, v \in V_c^q\}. \tag{23}$$

Event context may provide functional, action, causal, and temporal cues for retrieval and relation interpretation, but does not define an independent candidate node.

Before query-specific graph construction, semantic relations are restricted to transportable endpoints and deduplicated using the ordered triplet

$$(\text{subject}, \text{predicate}, \text{object}). \tag{24}$$

We distinguish two relation origins. An *evidence-linked relation* is retained from a visually grounded semantic annotation and preserves an explicit supporting image identifier. A *context-derived relation* is introduced conservatively from supported context within the same source unit. Implemented context-derived relation types include `shares_context`, `discusses_with`, and `located_in`. No synthetic relation is introduced across unrelated source units without explicit relational evidence.

Relation direction is preserved throughout graph construction because reversing the participants may change the narrative meaning of an interaction. A missing edge indicates lack of retained relational evidence rather than cultural incompatibility or impossibility.

For visual grounding, transportable records are linked to representative evidence images, and multiple records may share the same image. The Chinese VCM stores all 552 images as supporting evidence and exposes 276 representative images as inference-time references. The Persianate VCM exposes all 130 images as inference-time references.

Detailed triggering conditions, confidence assignments, treatment of missing relations, and topology normalization are provided in Appendix I.1.

Table 17: Full transportable-record graph statistics before query-specific retrieval. Evidence-linked edges originate from visually grounded semantic relation annotations after endpoint filtering and deduplication. Context-derived edges are constructed conservatively from supported information within the same source unit.

Cultural memory	Reference images	Transportable records	Evidence-linked edges	Context-derived edges	Total edges
Chinese	276	848	572	1,840	2,412
Persianate	130	520	650	1,170	1,820

J.5 Relation Confidence and Structural Usage

Each retained candidate-level relation is associated with a confidence $c_e \in [0, 1]$ determined by its evidence source and relation type. Confidence contributes to structural matching through

$$D_{\text{conf}}(e) = 1 - c_e, \tag{25}$$

which contributes $0.10(1-c_e)$ to the implemented relation discrepancy. Higher-confidence relations therefore incur smaller evidence penalties, whereas lower-confidence relations contribute larger discrepancies.

Relations lacking retained visual evidence or contextual support are excluded from the candidate graph rather than treated as supported cultural facts. Their absence indicates a gap in the available relational evidence, not cultural impossibility. The structural objective consequently uses finite missing-relation penalties rather than hard incompatibility constraints. Exact confidence assignments and missing-relation penalties are reported in Appendix I.1.

J.6 Separation from Evaluation Evidence

The generation-time VCMs are maintained separately from the evaluation-only cultural evidence corpus. VCM content is accessed during candidate retrieval, structural correspondence, reference selection, and visual realization. Evaluation evidence is not exposed during VCM construction or generation, and exact or near-duplicate generation-memory images are excluded from the evaluation evidence pool.

The Chinese and Persianate instantiations use the same narrative representation, memory schema, retrieval procedure, transport objective, solver, decoding procedure, and hyperparameters. The target context is changed by substituting the corresponding VCM, without culture-specific parameter adaptation or manually specified source-to-target mappings.

K GPT-4o Joint Assignment Baseline

We construct a GPT-4o Joint Assignment baseline to test whether the explicit structural correspondence in CultureWeaver can be replaced by general-purpose LLM reasoning under matched assignment inputs. GPT-4o receives the same narrative graph, candidate graph, candidate attributes, relational evidence, and relation confidence available to RA-srFGW. No additional retrieval or manually specified source-to-target mapping is provided. The downstream reference selection and rendering pipeline are kept fixed.

Assignment Input. For each transportable narrative node, the input includes its record type, role, narrative function, required affordances, incident relations, and narrative-importance score. Each cultural candidate includes its record type, semantic description, functional attributes, contextual information, and supported directed relations with confidence values. Event context is provided through the associated action, causal, and temporal information.

System Prompt. GPT-4o is instructed to solve the assignments jointly rather than matching narrative nodes independently:

> You are given a source narrative graph and a target-cultural candidate graph. Assign exactly one target-cultural candidate to each transportable narrative node.

Consider all assignments jointly. Each selected candidate should satisfy the source node's role, function, affordance, and type constraints, while the complete assignment should support the required directed relations among source nodes.

Use only the candidates and relational evidence provided in the input. Do not invent candidates, relations, or cultural facts. Treat a missing candidate relation as lack of supporting evidence rather than proof of cultural impossibility.

Distinct source-character roles must not share the same target-character identity.

Return only valid JSON following the required output schema.

Input Template. Each scene is represented using the following structured template:

```
SOURCE NARRATIVE REQUIREMENTS

Narrative nodes:
- node_id: <id>
  type: <character/object/scene>
  role: <role>
  function: <function>
  affordances: <required affordances>
  importance: <importance score>

Directed narrative relations:
- subject: <node_id>
  predicate: <relation>
  object: <node_id>
  causal/temporal context: <optional qualifiers>

Event context:
- <action, causal, and temporal description>


TARGET-CULTURAL CANDIDATES

Candidate records:
- candidate_id: <id>
  type: <character/object/scene>
  description: <semantic description>
  functions: <functional attributes>
  affordances: <affordances>
  context: <contextual attributes>

Supported candidate relations:
- subject: <candidate_id>
  predicate: <relation>
  object: <candidate_id>
  confidence: <relation confidence>


TASK

Assign exactly one candidate to each narrative node.
Consider both node-level functional compatibility and the directed relations
required by the narrative.
Use only the provided candidates and evidence.
Return the result using the required JSON schema.
```

Output Schema. GPT-4o returns one discrete cultural binding for each transportable narrative node:

```
{
  "assignments": [
    {
      "narrative_node_id": "...",
      "candidate_id": "..."
    }
  ]
}
```

The output is accepted only if every transportable narrative node receives one valid candidate and all character-identity constraints are satisfied. Unknown identifiers, missing assignments, identity violations, or invalid JSON trigger a retry with the same input.

Inference Settings. We use `gpt-4o` for all scenes with temperature 0, top-$p = 1$, and a maximum output length of 1,024 tokens. One assignment is sampled per scene. Invalid outputs are retried up to three times using the same prompt and structured input. No external tools or additional retrieval are available to the model.

Matched Rendering. After a valid binding is obtained, GPT-4o Joint Assignment uses the same reference-selection, canonical-frame generation, and image-to-video rendering procedure as RA-srFGW. Thus, the comparison changes only the assignment mechanism while keeping the candidate support, cultural evidence, and downstream rendering settings matched.