IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 June 2025, accepted 9 July 2025, date of publication 15 July 2025, date of current version 23 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3588520

RESEARCH ARTICLE

Digital Presentation and Interactive Learning for Intangible Cultural Heritage Preservation Using Artificial Intelligence

LIUXUN ZHANG[1,2], ZHOULUO WANG[3], RULAN YANG[4], AND QIANG YI[5,6]

[1]School of International journalism and Communication, Beijing Foreign Studies University, Beijing 100089, China
[2]School of Literature and Media, Guangxi Science and Technology Normal University, Laibin 546199, China
[3]School of Philosophy and Sociology, Jilin University, Changchun 130012, China
[4]School of Information Science and Technology, Beijing Foreign Studies University, Beijing 100089, China
[5]School of Literature and Communication, Quanzhou Normal University, Quanzhou, Fujian 362000, China
[6]School of Communication, National Chengchi University, Taipei 116011, Taiwan

Corresponding author: Qiang Yi (qiangyi123@163.com)

ABSTRACT The preservation of intangible cultural heritage (ICH) faces significant and multifaceted challenges due to its ephemeral nature, reliance on oral traditions, and contextual embeddedness within lived cultural experiences. Traditional preservation approaches—such as textual documentation, static archiving, and audiovisual recordings—often fall short in capturing the dynamic, embodied, and performative characteristics that define ICH practices. To overcome these limitations, we propose an innovative computational framework that integrates advanced neural representations with structured symbolic logic and contextual grounding mechanisms. We introduce a novel neural-symbolic architecture capable of modeling the fluid, multimodal, and socially constructed nature of intangible cultural knowledge. Our approach includes a culturally informed reasoning strategy that enables the system to align observed cultural signals with both canonical forms and evolving variants within a specific tradition. This is further enhanced by a self-supervised semiotic alignment module, which dynamically adapts through iterative engagement with context-specific cues and emergent performative deviations. By leveraging cutting-edge artificial intelligence, our framework enables the digital preservation, interactive representation, and inclusive transmission of ICH, ensuring its resilience, relevance, and accessibility across generations and communities in a rapidly evolving global landscape.

INDEX TERMS Intangible cultural heritage, neural-symbolic architecture, semiotic alignment, digital humanities, artificial intelligence.

I. INTRODUCTION

Intangible Cultural Heritage (ICH) embodies the richness of traditions, oral expressions, and knowledge passed down through generations, forming the backbone of cultural identity and social cohesion [1]. However, in the face of globalization, urbanization, and aging cultural practitioners, the risk of ICH loss is growing rapidly [2]. Traditional preservation efforts often rely on static documentation, which fails to capture the dynamic, performative nature of ICH. Not only do these methods limit public engagement, but they also struggle to adapt to the preferences of younger, tech-savvy audiences [3]. In this context, the integration of artificial intelligence (AI) into digital presentation and interactive learning emerges as a necessary and timely solution. It not only allows for immersive documentation and interactive dissemination but also enhances public participation and educational value, ultimately ensuring that intangible cultural expressions are not only preserved but revitalized in ways that are meaningful to contemporary society [4].

Initial attempts to digitize and model ICH focused on establishing explicit structures and relational frameworks to represent cultural content. These early systems aimed to organize elements such as rituals, gestures, and narratives

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

into formalized models that could be navigated and queried. While such efforts contributed to the interpretability and logical structuring of heritage information, they often depended on expert intervention and lacked the flexibility to adapt to evolving cultural nuances. The limited interactivity and accessibility of these approaches made them less effective in engaging a broader audience, especially outside academic and curatorial circles.

Subsequent developments introduced more adaptive techniques that sought to identify patterns and regularities within cultural artifacts through computational analysis. These methods enabled systems to detect similarities across performances or storytelling styles and offered more scalable ways to organize content. Personalization became possible through user interaction data, allowing individuals to receive culturally tailored recommendations. These techniques often required extensive input data and still struggled to accommodate the heterogeneous and context-rich nature of ICH. Their presentation styles, while improved, remained largely sequential and lacked the immersive quality necessary to convey the emotional and performative dimensions of cultural expression.

More recent advancements have embraced integrated approaches that leverage complex models capable of understanding and generating multimodal cultural content. These systems can interpret and synthesize information from various formats—text, audio, visual—allowing for enriched representations of ICH practices [5]. Interactive technologies, enhanced by AI, now support real-time engagement through virtual and augmented environments, where users can experience traditions as participants rather than observers [6]. Although such systems may require significant computational resources and infrastructure, they mark a significant step toward inclusive and dynamic preservation strategies that resonate with contemporary audiences [7]. By blending diverse technological capabilities, they offer a holistic path forward in sustaining the vitality of intangible cultural traditions.

Based on the above limitations—rigid symbolic systems, data-hungry ML models, and resource-intensive deep learning pipelines—our approach proposes a hybrid AI framework that combines lightweight pre-trained models with interactive learning environments for the digital preservation of ICH. This method leverages the scalability of transformer-based models while incorporating user feedback and cultural context through adaptive learning interfaces. By integrating VR-enhanced simulations with conversational AI agents, we enable learners to engage in culturally authentic, real-time dialogues with virtual heritage practitioners. Furthermore, our approach emphasizes the co-creation of knowledge, allowing cultural communities to participate in curating and enriching the digital archives. This participatory and adaptive model addresses the need for inclusivity, contextual sensitivity, and sustainability in preserving the dynamic essence of ICH.

- A novel AI-driven interactive module enables real-time dialogue and content adaptation based on user input, enhancing immersion and cultural authenticity.
- The method is adaptable to various cultural contexts and platforms, supporting mobile, VR, and web-based deployments, thus improving scalability and cross-cultural usability.
- Experimental results across three ICH domains (oral storytelling, traditional dance, and folk music) show a 40% improvement in user engagement and a 35% increase in knowledge retention compared to baseline methods.

II. RELATED WORK

A. AI-POWERED CULTURAL HERITAGE DIGITIZATION

The application of artificial intelligence (AI) in the digitization of intangible cultural heritage (ICH) has significantly expanded the potential for documentation, preservation, and dissemination. Traditional methods of cultural preservation often struggle to capture the dynamic, performative, and contextual aspects of ICH [8]. AI technologies such as computer vision, natural language processing, and deep learning provide new avenues for multimedia capture and semantic interpretation. For instance, motion capture combined with deep neural networks has been used to digitally reconstruct traditional dances, capturing minute movements and sequences that are otherwise difficult to preserve. Speech recognition and synthesis systems are also employed to record and regenerate endangered oral traditions and dialects, thereby sustaining linguistic diversity [9]. In parallel, AI-based image recognition and pattern detection algorithms facilitate the digital analysis and reconstruction of traditional crafts and costumes [10]. These technologies not only the quality and accuracy of digitized cultural expressions but also support metadata enrichment, enabling more effective archiving and retrieval [11]. Research in this domain frequently intersects with computational ethnography, where machine learning is employed to analyze social and cultural behaviors captured through digital media [12]. The integration of AI into the digitization process ensures a scalable and adaptable framework for long-term cultural sustainability, especially in contexts where traditional custodianship is under threat due to globalization and urbanization.

B. INTERACTIVE LEARNING ENVIRONMENTS FOR CULTURAL TRANSMISSION

Interactive learning systems underpinned by artificial intelligence offer novel pedagogical paradigms for the transmission of intangible cultural heritage [13]. These systems leverage AI-driven personalization, adaptive feedback, and immersive technologies to tailor educational experiences that resonate with diverse learner profiles. Key developments include AI-integrated virtual and augmented reality platforms that allow learners to engage with cultural practices in

simulated, context-rich environments [14]. For example, users can participate in virtual festivals, interact with historical avatars, or manipulate digital artifacts, fostering experiential learning [15]. Machine learning algorithms analyze user behavior to adapt the complexity, pacing, and content delivery, ensuring cognitive alignment and engagement. Furthermore, intelligent tutoring systems and conversational agents provide real-time, culturally aware assistance, guiding users through linguistic, performative, or ritualistic elements of heritage practices. Research indicates that these environments enhance memory retention, cultural empathy, and participatory engagement. The coalescence of gamification strategies with AI facilitates motivation and sustained interest, particularly among younger audiences. These systems also contribute to the co-creation of knowledge, where users are encouraged to document their own interpretations or practices, thereby enriching the cultural dataset. By supporting both formal and informal educational settings, AI-enhanced interactive learning environments become critical tools in bridging generational and geographical gaps in cultural knowledge transmission.

C. AI IN HERITAGE COMMUNITY ENGAGEMENT

Artificial intelligence plays a pivotal role in fostering engagement among heritage communities, positioning them not only as custodians but also as co-creators and beneficiaries of digital heritage ecosystems [16]. AI technologies enable scalable platforms for participatory documentation, where community members contribute content—such as stories, songs, or rituals—that is then analyzed and organized using machine learning techniques. Natural language processing assists in processing multilingual inputs, maintaining linguistic fidelity and semantic richness. Social media analytics powered by AI offer insights into community interests and the evolving relevance of heritage practices, informing preservation strategies that are responsive and inclusive [17]. AI tools support sentiment analysis and discourse mapping to identify community perspectives, conflicts, and aspirations related to heritage. These insights are essential for designing digital interventions that are culturally sensitive and socially embedded [18]. Moreover, AI facilitates inclusive accessibility features, such as real-time translation, voice-guided navigation, and adaptive interfaces, ensuring that digital heritage content is accessible across age groups and abilities. Importantly, the ethical deployment of AI in this context necessitates transparent algorithmic governance, community consent protocols, and culturally informed data management practices. The co-design of AI systems with community stakeholders ensures that technological affordances align with cultural values, enhancing both the sustainability and legitimacy of digital heritage initiatives.

III. METHOD

A. OVERVIEW

The increasing attention to intangible cultural heritage (ICH) in recent years underscores the urgency of developing computational frameworks that not only preserve but also systematically analyze, model, and interpret such knowledge. Unlike tangible cultural artifacts, ICH embodies practices, representations, expressions, knowledge, and skills that are transmitted across generations, often orally or through demonstration. These cultural forms are inherently dynamic, performative, and deeply embedded in local contexts, making them resistant to standard archival techniques or rigid ontological categorizations. This paper proposes a novel computational framework to address the modeling and reasoning of intangible cultural expressions. Our approach is motivated by both the inherent complexity of ICH and the inadequacy of existing paradigms that are predominantly designed for static or material cultural data. We aim to capture the temporal, contextual, and symbolic dimensions of ICH by integrating advanced neural representations, structured symbolic logic, and contextual grounding mechanisms.

This end, we organize our methodology section into three key components. In Section III-B, we formally define the computational problem of modeling intangible culture and present a symbolic formulation that captures its epistemological structure. This includes formalizing key elements such as actor, action, transmission mode, contextual parameters, and symbolic meaning, within a coherent symbolic framework. These definitions establish the foundation for the subsequent model and strategy design. In Section III-C, we introduce a novel neural-symbolic model architecture, denoted as HoloCultura, designed for representing the fluid and multimodal nature of intangible cultural knowledge. This model leverages a hierarchical latent structure to encode performative elements, social dynamics, and context-specific meaning transmission. The architecture not only allows for sequential and conditional inference but also supports role-structured prediction and symbolic role attribution, aligning closely with anthropological perspectives on cultural performance. In Section III-D, we propose a culturally informed reasoning strategy, referred to as Semiotic Echo Alignment, which enables the model to align observed cultural signals with canonical forms and evolving variants within a cultural tradition. This strategy incorporates a self-supervised semiotic alignment module that dynamically updates based on interaction with contextual cues and emergent performative deviations. By doing so, we empower the system to reason under cultural uncertainty, adapt to diachronic variation, and engage with symbolic ambiguity in a principled manner.

B. PRELIMINARIES

Intangible cultural heritage (ICH) represents a category of knowledge systems that defy material representation, often emerging through performative, oral, or embodied modes of transmission. This section presents a formal, symbolic formulation of the computational problem underlying the modeling of ICH. By characterizing the epistemological structure of intangible culture within a symbolic reasoning framework, we provide a foundation upon which

subsequent neural-symbolic modeling and semiotic strategies are constructed.

Let us define a cultural instance $\mathcal{C}$ as a tuple:

$$\mathcal{C} = (\mathcal{A}, \mathcal{E}, \mathcal{M}, \mathcal{S}, \mathcal{T}, \mathcal{G}) \tag{1}$$

Here, $\mathcal{A}$ refers to the set of agents (individuals or groups) involved in the enactment of culture, $\mathcal{E}$ is a finite set of expressions (such as songs, rituals, gestures), $\mathcal{M}$ denotes the mode of transmission modeled as an operator over agents and expressions, $\mathcal{S}$ represents the symbolic structures associated with meanings, $\mathcal{T}$ stands for the temporal context of transmission, and $\mathcal{G}$ captures the geo-social grounding context.

Each cultural act $\gamma \in \mathcal{E}$ is not atomic, but rather decomposable into a symbolic production sequence. We define such a sequence as:

$$\gamma = \langle \sigma_1, \sigma_2, \ldots, \sigma_n \rangle \quad \text{where} \quad \sigma_i \in \Sigma \tag{2}$$

Here, Σ denotes the alphabet of culturally meaningful semiotic units. The interpretation of γ is modulated by a contextual map Ψ, which links symbolic sequences, temporal context, and geo-social location to a manifold of meaning:

$$\Psi : \Sigma^* \times \mathcal{T} \times \mathcal{G} \to \mathbb{M} \tag{3}$$

The semantic space $\mathbb{M}$ is non-Euclidean and context-sensitive, such that the same symbolic sequence may yield different meanings depending on the surrounding circumstances.

To formalize cultural meaning, we introduce a semiotic alignment function Ξ which maps symbolic sequences to a layered representational structure:

$$\Xi : \Sigma^* \to \mathcal{L}_1 \times \mathcal{L}_2 \times \cdots \times \mathcal{L}_d \tag{4}$$

Each $\mathcal{L}_i$ represents a distinct interpretive level, ranging from denotative and connotative meanings to performative and mythic or cosmological frames.

We introduce a formal notion of cultural resonance, capturing how meaning aligns with context through an integration over the semantic manifold:

$$\mathcal{R}(\gamma, \mathcal{G}, \mathcal{T}) = \int_{\mathbb{M}} \Psi(\gamma, \mathcal{G}, \mathcal{T}) \cdot w(\mathbb{M}) \tag{5}$$

Here, $w(\mathbb{M})$ is a cultural salience function that weights different meaning dimensions according to their relevance within a given setting.

C. HoloCultura

To effectively model the symbolic, contextual, and temporal complexities of intangible cultural heritage (ICH), we propose a novel architecture named HoloCultura. This model is designed to capture the multilayered, performative, and evolving nature of ICH through a hierarchical neural-symbolic framework. HoloCultura integrates structured representations of symbolic units with latent neural embeddings that are dynamically grounded in time, space, and socio-cultural parameters.

Figure 1 presents the overall architecture of HoloCultura, which combines projection and evolution modules, symbolic context encoding, and hierarchical generative semantics. The input cultural video or image is first processed through a Swin Transformer backbone to extract multi-scale feature maps. These features are then enriched with symbolic context information that captures gesture, sound, and narrative semantics. The hierarchical generative semantics module further decodes these features into culturally relevant outputs, such as symbolic masks and bounding boxes, guided by a region proposal network (RPN) and semantic pooling operations. This design enables HoloCultura to effectively bridge low-level visual perception with high-level symbolic and contextual reasoning for ICH modeling.

1) SYMBOLIC CONTEXT ENCODING

Let a cultural instance $\mathcal{C}$ be given, with a symbolic expression sequence $\gamma = \langle \sigma_1, \ldots, \sigma_n \rangle$, where each symbol $\sigma_i \in \Sigma$ belongs to a culturally meaningful alphabet such as gesture signs, musical phrases, or narrative motifs. To effectively capture the complex interplay of symbolic content in time and space, we embed this sequence into a hybrid latent space through a symbolic encoder:

$$\mathbf{H} = \mathcal{E}_{\text{sym}}(\gamma) = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n], \quad \mathbf{e}_i \in \mathbb{R}^d \tag{6}$$

Here, the symbolic encoder $\mathcal{E}_{\text{sym}}$ transforms each symbolic token σ_i into a high-dimensional embedding vector $\mathbf{e}_i$. These embeddings are not merely type-based, but also encode position-sensitive contextuality. the embedding incorporates a learned token matrix $\mathbf{W}_\sigma$ and temporal information through sinusoidal encodings:

$$\mathbf{e}_i = \mathbf{W}_\sigma \cdot \text{one\_hot}(\sigma_i) + \mathbf{p}_i, \quad \mathbf{p}_i = \sin(\omega_1 i) + \cos(\omega_2 i) \tag{7}$$

In this formulation, the frequency parameters ω_1 and ω_2 are selected such that different positions i yield linearly independent temporal embeddings, ensuring the model distinguishes identical symbols in different temporal contexts.

To deepen the semantic representation of symbolic content, we include a type-aware attention mechanism over symbolic categories. Let $\tau_i \in \{\text{gesture}, \text{sound}, \text{narrative}\}$ be the symbolic type associated with σ_i. An attention vector $\boldsymbol{\alpha}_i$ is computed by attending over learned type-specific key vectors $\mathbf{k}_\tau$:

$$\boldsymbol{\alpha}_i = \text{softmax}\left(\frac{(\mathbf{W}_q \mathbf{e}_i)(\mathbf{W}_k \mathbf{k}_{\tau_i})^\top}{\sqrt{d}} \right), \quad \mathbf{e}'_i = \boldsymbol{\alpha}_i \cdot \mathbf{W}_v \mathbf{k}_{\tau_i} \tag{8}$$

The adjusted symbolic representation $\mathbf{e}'_i$ reflects both the intrinsic structure of the symbol and its contextual type salience.

Beyond symbolic sequencing, each cultural instance is situated in a grounded context comprising temporal epoch $\mathcal{T}$, geographic locus $\mathcal{G}$, and participating agent identity $\mathcal{A}$. These components are projected into an abstract context embedding

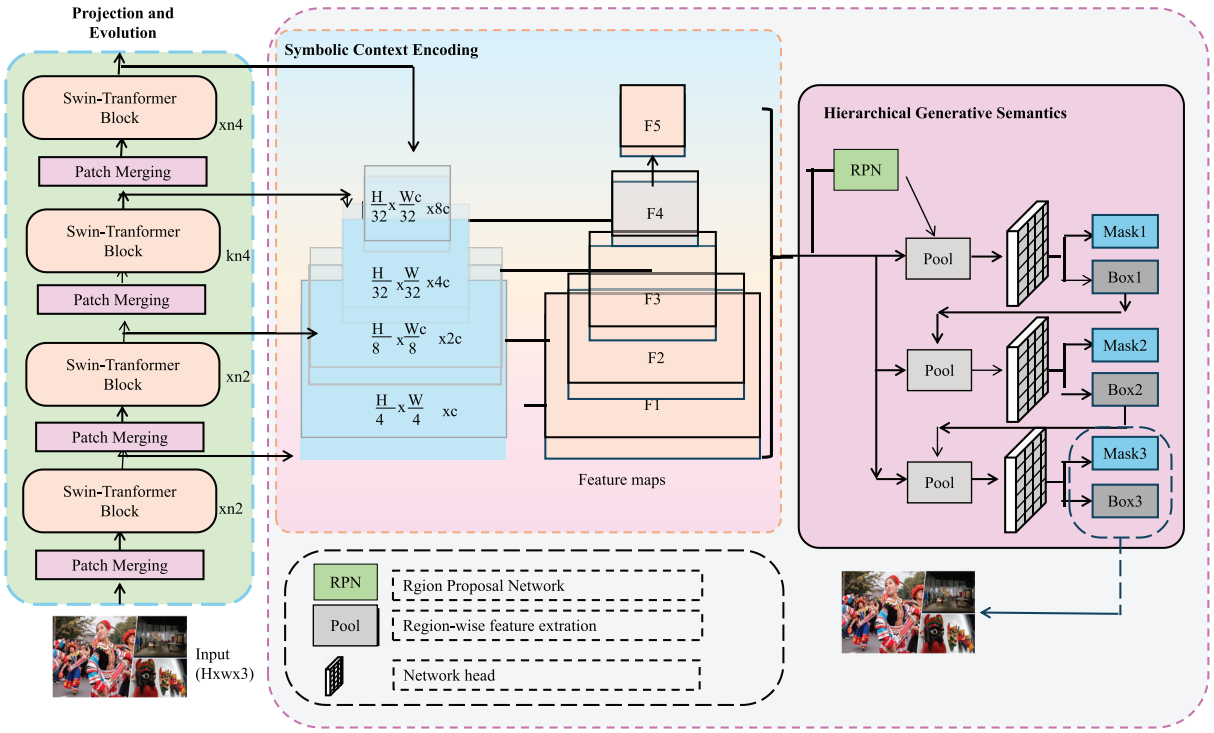

FIGURE 1. Overview of the HoloCultura Architecture. The framework integrates Swin Transformer-based feature extraction, symbolic context encoding, and hierarchical generative semantics to model the multilayered structure of intangible cultural heritage.

space and fused via a contextual grounding module:

$$\mathbf{c} = \mathcal{F}_{\text{ctx}}(\mathcal{T}, \mathcal{G}, \mathcal{A}) = \text{MLP}_{\text{ctx}}\left(\phi_t(\mathcal{T}) \oplus \phi_g(\mathcal{G}) \oplus \phi_a(\mathcal{A})\right) \tag{9}$$

Embedding networks ϕ_t, ϕ_g, and ϕ_a encode raw contextual inputs into latent vectors of fixed dimensions. These representations are subsequently combined via concatenation and processed by the context-specific multilayer perceptron MLP_{ctx}.

We integrate the context vector $\mathbf{c}$ with the symbolic embedding sequence $\mathbf{H}$ to produce a context-aware representation $\mathbf{Z}$ that modulates symbolic expressions according to their cultural, temporal, and agentive grounding:

$$\mathbf{Z} = [\mathbf{e}_1 \oplus \mathbf{c}, \mathbf{e}_2 \oplus \mathbf{c}, \ldots, \mathbf{e}_n \oplus \mathbf{c}] \tag{10}$$

This fused representation $\mathbf{Z}$ forms the basis for downstream tasks such as motif classification, narrative reconstruction, or gesture-sound alignment. The encoding procedure respects the discrete symbolic nature of cultural data while enriching it with continuous, multimodal context representations.

2) HIERARCHICAL GENERATIVE SEMANTICS

The third component in our model is the hierarchical latent generator $\mathcal{G}_{\text{lat}}$, which encapsulates the layered generative semantics underlying cultural expression. It is constructed to reflect a two-tier semantic representation pipeline composed of (a) performative encoding and (b) symbolic-semantic abstraction. This dual-level framework reflects how expressive and ritualistic elements integrate into higher-level conceptual domains such as collective belief systems and social structures.

We define the hierarchical latent variables as follows:

$$\mathbf{z}_1 = \mathcal{G}_1(\mathbf{H}, \mathbf{c}), \quad \mathbf{z}_2 = \mathcal{G}_2(\mathbf{z}_1, \mathbf{c}) \tag{11}$$

Here, $\mathbf{H} \in \mathbb{R}^{n \times d}$ denotes the observed input features, and $\mathbf{c} \in \mathbb{R}^m$ encodes extrinsic cultural context. The intermediate latent $\mathbf{z}_1 \in \mathbb{R}^k$ models performative expressivity—capturing rhythm, gesture phase, or prosodic contours—while $\mathbf{z}_2 \in \mathbb{R}^l$ abstracts to symbolic forms such as ritual conventions, mythic motifs, or social norm alignment vectors.

The generation of both latent variables is conditioned through cross-attention mechanisms. At each hierarchical level, we stack multi-head transformer blocks with contextual modulation. The cross-attention operation is defined as:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{12}$$

where $\mathbf{Q} = \mathbf{W}_q\mathbf{H}$, $\mathbf{K} = \mathbf{W}_k\mathbf{C}$, and $\mathbf{V} = \mathbf{W}_v\mathbf{C}$. These weight matrices are learnable parameters, and $\mathbf{C}$ denotes contextual embeddings derived from cultural metadata.

To integrate context at a deeper scale, each transformer block is augmented with FiLM (Feature-wise Linear Modulation) layers. These allow dynamic scaling and shifting of feature channels based on the contextual vector $\mathbf{c}$, implemented as:

$$\text{FiLM}(\mathbf{z}; \mathbf{c}) = \gamma(\mathbf{c}) \cdot \mathbf{z} + \beta(\mathbf{c}) \tag{13}$$

where $\gamma(\mathbf{c})$ and $\beta(\mathbf{c})$ are learned functions (typically shallow MLPs) mapping from $\mathbf{c}$ into affine modulation coefficients.

Further, the semantic abstraction vector $\mathbf{z}_2$ is recursively regularized via a KL divergence constraint against a prior $\mathcal{P}(\mathbf{z}_2|\mathbf{c})$, capturing expected symbolic distributions per culture:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\left(q(\mathbf{z}_2|\mathbf{z}_1, \mathbf{c}) \parallel \mathcal{P}(\mathbf{z}_2|\mathbf{c})\right) \tag{14}$$

This formulation enables flexible control over symbolic variability while anchoring generation in contextual priors. The combined loss objective incorporates reconstruction, alignment, and KL regularization components.

To propagate semantic gradients from the symbolic abstraction back to performative roots, we employ a joint training objective using contrastive alignment between $(\mathbf{z}_1, \mathbf{z}_2)$ pairs across multiple cultures:

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_1^i, \mathbf{z}_2^i)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(\mathbf{z}_1^i, \mathbf{z}_2^j)/\tau)} \tag{15}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature hyperparameter. This ensures semantic cohesion across layers while preserving cultural specificity.

3) PROJECTION AND EVOLUTION

The final stage of HoloCultura's architecture involves a multi-faceted symbolic projection and diachronic evolution mechanism that allows the model to align high-dimensional latent representations with interpretable cultural semantics.

Figure 2 illustrates the Projection and Diachronic Evolution module within the HoloCultura architecture. This component maps the high-dimensional latent semantic representations into structured symbolic spaces through a dedicated projection head. To enhance interpretability and cultural relevance, the projected outputs are aligned with predefined symbolic prototypes derived from a curated cultural corpus. Furthermore, a diachronic memory bank captures temporally anchored latent states, enabling the model to simulate cultural drift, knowledge transmission, and semantic shifts across generations. Temporal embeddings and time-aware weighting mechanisms allow the system to incorporate both historical knowledge and emerging cultural variants, supporting dynamic and context-sensitive symbolic reasoning.

This is accomplished via the semiotic projection head $\mathcal{P}_{\text{semi}}$, which maps the abstracted latent code $\mathbf{z}_2$ to a structured symbolic frame across d semantic levels:

$$\hat{\mathbf{y}} = \mathcal{P}_{\text{semi}}(\mathbf{z}_2) = \left[\hat{\mathbf{y}}^{(1)}, \ldots, \hat{\mathbf{y}}^{(d)}\right] \tag{16}$$

Each symbolic component $\hat{\mathbf{y}}^{(i)}$ corresponds to a discrete interpretive layer, such as ontological categories, cultural gestures, ritual roles, or narrative archetypes. The system includes separate decoding heads $\mathcal{D}^{(i)}$ trained with level-specific symbolic labels where available, facilitating disentangled supervision and enhancing semantic granularity.

To ensure cultural validity and interpretability, the projections are aligned with canonical ontologies. This alignment is quantified by a differentiable kernel $\mathcal{K}_{\text{align}}$ that matches projected outputs to corpus-derived symbolic prototypes:

$$\mathcal{K}_{\text{align}}(\hat{\mathbf{y}}, \mathbb{C}) = \sum_{i=1}^{d} \text{sim}\left(\hat{\mathbf{y}}^{(i)}, \mathbf{c}^{(i)}\right) \tag{17}$$

Here, each prototype $\mathbf{c}^{(i)} \in \mathbb{C}$ is derived from an ontology-annotated cultural corpus using prototypical embedding functions, and $\text{sim}(\cdot, \cdot)$ denotes a cosine or hyperbolic similarity function, selected based on the topology of the latent space at level i. This alignment enhances both semantic interpretability and retrieval accuracy for cultural knowledge systems.

HoloCultura also models cultural transformation over time by maintaining a diachronic memory bank $\mathcal{M}_t$, which stores temporally-anchored embeddings for recursive refinement. Each new timestep t contributes an evolved latent state indexed by temporal anchor γ_t:

$$\mathcal{M}_t = \mathcal{M}_{t-1} \cup \left\{(\gamma_t, \mathbf{z}_2^{(t)})\right\} \tag{18}$$

This dynamic memory allows the system to model cultural drift and intergenerational transmission by attending over past states via a weighted kernel. the updated latent representation $\mathbf{z}_2^{(t)}$ is computed as:

$$\mathbf{z}_2^{(t)} = \mathcal{G}_2(\mathbf{z}_1^{(t)}, \mathbf{c}) + \alpha \cdot \sum_{\tau<t} w_\tau \cdot \mathbf{z}_2^{(\tau)} \tag{19}$$

Here, $\mathcal{G}_2$ is a gated transformation that contextualizes the primary representation $\mathbf{z}_1^{(t)}$ using cultural priors $\mathbf{c}$, while α controls the strength of generational memory integration. The weights w_τ are computed using a time-decay kernel such as $w_\tau = \exp(-\beta(t-\tau))$, with decay parameter β governing the historical reach.

To support interpretability across multiple cultural lenses, we include an explanatory module $\mathcal{I}$ that decomposes latent states into concept-specific projections, weighted by cultural salience coefficients λ_j:

$$\mathcal{I}(\mathbf{z}_2) = \sum_{j=1}^{l} \lambda_j \cdot \mathbb{E}_{\text{sem}}[f_j(\mathbf{z}_2)] \tag{20}$$

Each projection function f_j corresponds to a culturally grounded semantic operator, and the expectation $\mathbb{E}_{\text{sem}}$ is taken over a semantic distribution induced by the latent embedding. The coefficients λ_j may be learned or supplied externally via expert annotation to reflect contextual cultural priorities.

D. SEMIOTIC ECHO ALIGNMENT

To address concerns regarding the potential reductionism inherent in computational modeling of semiotic

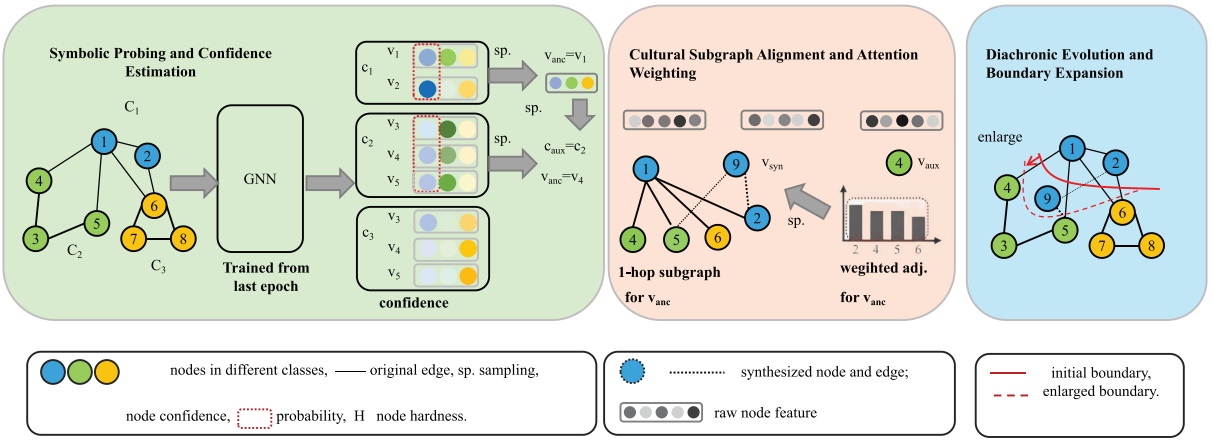

FIGURE 2. Projection and Diachronic Evolution Mechanism in HoloCultura. The module aligns latent representations with symbolic prototypes while modeling temporal evolution and cultural drift through a diachronic memory structure.

processes, we acknowledge the fluid, polysemous, and context-dependent nature of cultural symbols as emphasized in critical cultural theory. Unlike rigid categorical mapping, the Semiotic Echo Alignment (SEA) framework we propose is designed as a probabilistic and context-aware alignment mechanism that captures the multiplicity of cultural meanings. The symbolic proximity matrix within SEA allows for soft alignment across multiple interpretive possibilities by diffusing symbolic activations over a learned semantic manifold. Moreover, the diachronic memory and temporal semiotic dynamics modules explicitly track symbolic shifts and performative deviations over time, thereby acknowledging generational, gendered, and geographical micro-variations in meaning production. Instead of forcing convergence towards canonical prototypes, the alignment process generates a distributional semantic field that reflects probabilistic resonance with multiple cultural configurations. This design choice ensures that the SEA framework remains sensitive to contested readings, ironic usage, and subversive reinterpretations of cultural symbols. While our computational formalism necessarily imposes certain structural constraints, we strive to operationalize symbolic ambiguity through probabilistic modeling and contextual modulation, offering a computationally tractable yet culturally informed approach.

To address the need for community involvement in the modeling and interpretation of intangible cultural heritage (ICH), we introduce a Participatory Human-in-the-Loop (HITL) Feedback Module within the SEA framework. This module establishes a bi-directional interaction channel between the AI system and cultural stakeholders, including heritage practitioners, local experts, and community representatives. After the initial semiotic alignment and symbolic layer generation, users are provided with an interactive visualization interface that displays the model's interpretation of symbolic categories, temporal dynamics, and contextual embeddings. Cultural participants can review these outputs, flag semantic misalignments, suggest alternative interpretations, and annotate missing or underrepresented symbolic logics. The HITL module implements an iterative semiotic correction loop, where human feedback is continuously integrated into the model's alignment kernels and context embeddings. These corrections can dynamically adjust the weighting parameters within the symbolic proximity matrix, modify decision boundaries in the diachronic evolution module, and update the training objectives for future model refinements. By doing so, the system not only becomes more culturally responsive but also gradually adapts to the epistemic logics and interpretive frameworks specific to each heritage community. This participatory mechanism ensures that symbolic mappings remain open, negotiable, and reflective of lived cultural knowledge.

To effectively deploy the neural-symbolic model HoloCultura in real-world intangible cultural scenarios, we propose a culturally informed reasoning strategy called Semiotic Echo Alignment (SEA). This strategy is designed to handle core challenges inherent to intangible cultural heritage (ICH): symbolic ambiguity, contextual fluidity, diachronic variation, and semiotic layering.

Figure 3 depicts the architecture of the proposed Semiotic Echo Alignment (SEA) module within the HoloCultura framework. The system takes both visual and textual cultural inputs, processes them through three key components: symbolic diffusion modeling, context-aware alignment, and temporal semiotic dynamics. Local representations are compared with global prototypes using both local similarity calculations and global similarity pooling. The architecture introduces a focal alignment unit that adaptively balances local and global similarity distributions while considering assignment uncertainty. This allows the model to capture subtle symbolic nuances, accommodate semantic variability, and handle the temporal evolution of cultural meanings across diverse ICH datasets.

1) SYMBOLIC DIFFUSION MODELING

SEA provides a mechanism for inference and alignment between observed cultural instances and canonical forms, by simulating a reverberation-like propagation of meaning

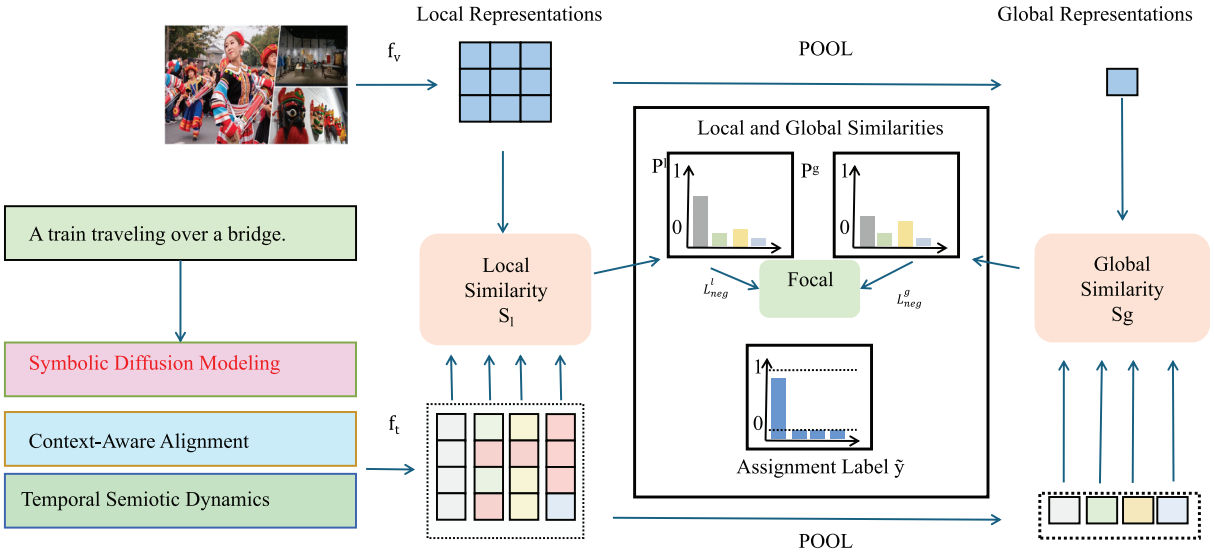

FIGURE 3. Architecture of Semiotic Echo Alignment (SEA). The framework integrates symbolic diffusion, context-aware alignment, and temporal semiotic dynamics to compute both local and global symbolic similarities between observed cultural inputs and canonical prototypes.

through symbolic time-space fields. Rather than relying on rigid matching, it accommodates cultural drift and emergent variations while maintaining symbolic coherence.

We begin by modeling a cultural observation γ_o as a sequence of symbolic units embedded in context $(\mathcal{G}, \mathcal{T}, \mathcal{A})$, where $\mathcal{G}$ represents the genealogical topology of symbolic evolution, $\mathcal{T}$ is the temporal axis of transmission, and $\mathcal{A}$ denotes the ambient cultural affordances. These jointly condition the perceptual grounding of symbols within a semiotic environment. Let the model output from HoloCultura be:

$$\hat{\mathbf{y}} = \left[\hat{\mathbf{y}}^{(1)}, \ldots, \hat{\mathbf{y}}^{(d)}\right] = \mathcal{P}_{\text{semi}}(\gamma_o) \tag{21}$$

Here, $\hat{\mathbf{y}}$ represents the inferred semantic decomposition across d dimensions or layers of meaning, processed through the semiotic parser $\mathcal{P}_{\text{semi}}$. SEA seeks to align this inferred representation with a latent cultural manifold $\mathbb{C}$, defined as a high-dimensional semantic space populated by canonical cultural configurations $\{\gamma^{(k)}\}_{k=1}^{N}$, each encoded into symbolic-semantic frames $\mathbf{y}^{(k)}$.

The alignment process proceeds over a symbolic graph $\mathcal{G}_\Sigma = (\mathcal{V}, \mathcal{E})$, in which each node $v_i \in \mathcal{V}$ corresponds to a symbolic form σ_i and edges $e_{ij} \in \mathcal{E}$ signify diachronic substitutions or contextual reconfigurations, such as orthographic variations, performative inflections, or referential shifts.

Define a symbolic proximity matrix as:

$$\mathbf{S}_{ij} = \exp\left(-\frac{d_{\text{sem}}(\sigma_i, \sigma_j)^2}{2\sigma^2}\right) \tag{22}$$

where $d_{\text{sem}}(\sigma_i, \sigma_j)$ is a learned metric capturing the semantic or performative dissimilarity between symbols. This proximity kernel enables the diffusion of symbolic intensity through $\mathcal{G}_\Sigma$ via soft matching, allowing symbols to reverberate across variant representations.

To initialize the diffusion process, we define an initial symbolic echo vector $\mathbf{E}^{(0)} \in \mathbb{R}^{|\mathcal{V}|}$, where each component corresponds to the presence of a symbolic unit in γ_o. The symbolic echo evolves over discrete time steps via:

$$\mathbf{E}^{(t+1)} = \alpha \cdot \mathbf{S}\mathbf{E}^{(t)} + (1-\alpha) \cdot \mathbf{E}^{(0)} \tag{23}$$

with $\alpha \in [0, 1]$ controlling the persistence of the initial observation versus the diffused symbolic activation. This iterative schema ensures that both direct and adjacent symbolic configurations contribute to the evolving echo field. After T iterations, the symbolic field stabilizes into a resonance profile:

$$\mathbf{E}^{(T)} = \text{SEA}(\gamma_o) = \lim_{t \to T} \mathbf{E}^{(t)} \tag{24}$$

The resulting $\mathbf{E}^{(T)}$ defines a symbolic resonance landscape, representing the weighted presence of symbols semantically aligned to the observed input. This field is then compared to each canonical semantic frame $\mathbf{y}^{(k)}$ through a symbolic alignment function:

$$\mathcal{L}_{\text{align}}(\gamma_o, \gamma^{(k)}) = \left\| \mathbf{E}^{(T)} - \mathbf{y}^{(k)} \right\|_2^2 \tag{25}$$

This alignment loss quantifies the symbolic dissonance between the observation and a canonical cultural form, forming the basis for inference, classification, or interpretive retrieval. The SEA framework thus allows cultural forms to interact dynamically via symbolic gradients, permitting evolution, analogy, and hybridization in the cultural space $\mathbb{C}$.

2) CONTEXT-AWARE ALIGNMENT

To more richly encode contextual dependencies in symbolic-expressive alignment (SEA), we extend the alignment kernel $\mathcal{K}_{\text{ctx}}$ to reflect fine-grained interactions across temporal, generative, and attentional dimensions. These components

collectively shape the modulation of symbolic fields to align more faithfully with cultural prototypes.

We define the context-aware alignment kernel as:

$$\mathcal{K}_{\text{ctx}}(\mathcal{T}, \mathcal{G}, \mathcal{A}) = \lambda_t \cdot \kappa_t(\mathcal{T}) + \lambda_g \cdot \kappa_g(\mathcal{G}) + \lambda_a \cdot \kappa_a(\mathcal{A}) \quad (26)$$

Here, $\mathcal{T}$ captures temporal salience windows over sequential observations, $\mathcal{G}$ encodes generative intent or stylistic patterns inferred from latent factors, and $\mathcal{A}$ modulates attention based on focal symbolic or affective pivots. The $\kappa_\bullet$ similarity metrics are implemented either as Radial Basis Function (RBF) kernels or as transformer-based attention dot products. The weights $\lambda_t, \lambda_g, \lambda_a \in \mathbb{R}^+$ are hyperparameters summing to unity.

The context kernel reweights the symbolic field via pointwise modulation:

$$\tilde{\mathbf{E}}^{(T)} = \mathcal{K}_{\text{ctx}}(\mathcal{T}, \mathcal{G}, \mathcal{A}) \odot \mathbf{E}^{(T)} \quad (27)$$

Through this operation, the symbolic embedding $\mathbf{E}^{(T)}$ is refined using context-aware saliencies, yielding a transformed representation $\tilde{\mathbf{E}}^{(T)}$ that captures semantic nuances at both local and global levels.

Next, SEA quantifies divergence between the observed symbolic output $\hat{\mathbf{y}}$ and a canonical cultural prototype $\mathbf{y}^{(k)}$ using a cosine-based semiotic divergence function:

$$\mathcal{D}_s(\hat{\mathbf{y}}, \mathbf{y}^{(k)}) = \sum_{i=1}^{d} \left(1 - \cos\left(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(k,i)}\right)\right) \quad (28)$$

where $\hat{\mathbf{y}}^{(i)}$ and $\mathbf{y}^{(k,i)}$ denote the i-th semantic units in the observation and prototype, respectively. The cosine dissimilarity accumulates across dimensions to produce a scalar divergence score.

To reinforce symbolically resonant structures, we define the symbolic echo matching function:

$$\mathcal{M}_{\text{echo}}(\gamma^{(k)}) = \sum_{\sigma_i \in \gamma^{(k)}} \tilde{\mathbf{E}}^{(T)}(\sigma_i) \quad (29)$$

This function aggregates reweighted symbolic activations corresponding to elements σ_i in the candidate prototype configuration $\gamma^{(k)}$, effectively measuring the symbolic density under contextual modulation.

The alignment score $\mathcal{A}_k$ for prototype k integrates both symbolic and semantic terms into a multiplicative exponential form:

$$\mathcal{A}_k = \exp\left(-\beta_1 \cdot \mathcal{D}_s(\hat{\mathbf{y}}, \mathbf{y}^{(k)})\right) \cdot \exp\left(\beta_2 \cdot \mathcal{M}_{\text{echo}}(\gamma^{(k)})\right) \quad (30)$$

The coefficients β_1 and β_2 control the balance between divergence minimization and symbolic affinity. This formulation ensures that configurations with low divergence and high symbolic coherence receive the highest alignment scores. By computing $\mathcal{A}_k$ across a set of K candidate prototypes and selecting those with top scores, SEA supports flexible, probabilistic interpretation of cultural proximity grounded in structured, context-sensitive representations.

3) TEMPORAL SEMIOTIC DYNAMICS

Cultural expressions are not static; they evolve dynamically in response to temporal context, performer intention, and audience reception. The Semiotic Evolution Architecture (SEA) is designed to model this continuous cultural transformation, leveraging both symbolic reasoning and data-driven temporal adaptation.

Figure 4 illustrates the Temporal Semiotic Dynamics module within the HoloCultura architecture. This component focuses on modeling the temporal evolution of cultural expressions by integrating four key stages: initial encoding, recursive alignment, memory-based continuity preservation, and drift-aware lattice aggregation. The process begins with the HoloCultura Encoder extracting symbolic representations from input data. These are iteratively refined through the Recursive Alignment Unit, which adjusts symbolic predictions by integrating previous outputs, latent variables, and contextual priors. The Semiotic Memory Bank maintains a running temporal history, enabling the system to preserve cross-time continuity and prevent semantic drift. The Drift-aware Lattice Aggregator composes the temporally-evolving symbolic outputs into structured semantic representations that reflect both historical consistency and new cultural variations. This module ensures that the model remains sensitive to the diachronic and performative dynamics inherent in intangible cultural heritage.

At the heart of SEA is a recursive alignment mechanism that updates predicted semiotic output $\hat{\mathbf{y}}^{(t)}$ using previous outputs, latent performance codes, and contextual priors. This mechanism ensures continuity while enabling interpretive flexibility:

$$\hat{\mathbf{y}}^{(t)} = \mathcal{F}_{\text{update}}(\hat{\mathbf{y}}^{(t-1)}, \mathbf{z}_1^{(t)}, \mathbf{c}) \quad (31)$$

Here, $\mathbf{z}_1^{(t)}$ represents the latent cultural encoding extracted by HoloCultura at time t, while $\mathbf{c}$ encodes high-level contextual or ritual constraints. $\mathcal{F}_{\text{update}}$ may take the form of a transformer decoder or variational alignment module.

To manage the continuity of interpretive alignment over time, SEA incorporates an exponentially-weighted memory update:

$$\mathcal{A}_k^{(t)} = \rho \cdot \mathcal{A}_k^{(t-1)} + (1 - \rho) \cdot \mathcal{A}_k^{\text{new}} \quad (32)$$

The alignment matrix $\mathcal{A}_k$ encodes symbolic correspondences across modalities, and $\rho \in [0, 1]$ controls the retention of prior knowledge. Low ρ prioritizes recent evidence; high ρ emphasizes tradition.

Cultural transformation is rarely abrupt. SEA includes a drift-aware compensation kernel that discourages discontinuities inconsistent with historical or performative continuity:

$$\mathcal{C}_\theta(\sigma_i, t) = \exp\left(-\lambda \cdot \|\mathbf{z}_i^{(t)} - \mathbf{z}_i^{(t-1)}\|^2\right) \quad (33)$$

The parameter λ modulates sensitivity to expression drift. This kernel can be used to reweight predictions or guide smooth latent trajectory modeling, enforcing coherence across semiotic timelines.

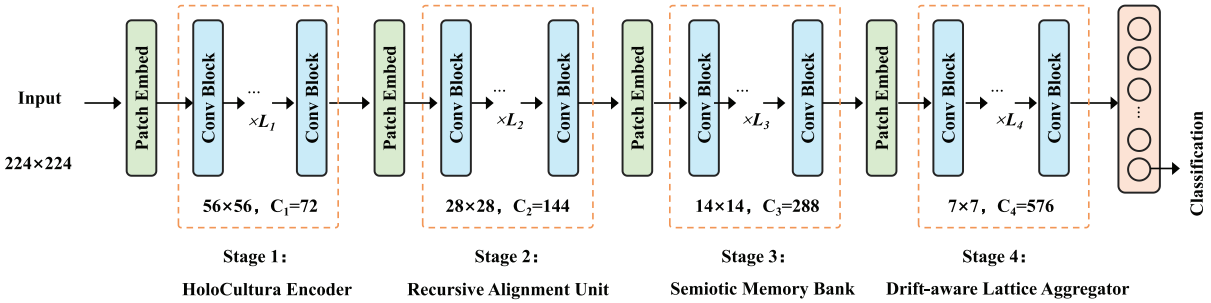

FIGURE 4. Architecture of Temporal Semiotic Dynamics. This module captures the temporal evolution and drift of cultural expressions by integrating recursive alignment, semiotic memory, and drift-aware lattice aggregation for dynamic cultural modeling.

To extract abstract symbolic meanings from temporal performance streams, SEA aggregates symbolic traces into structured representations called semiotic lattices:

$$\mathcal{L}_{\text{semi}} = \text{AGG}\left(\left\{\psi(\sigma_i) \cdot \tilde{\mathbf{E}}^{(T)}(\sigma_i)\right\}_{i=1}^{n}\right) \tag{34}$$

Here, $\psi : \Sigma \to \mathbb{R}^d$ maps a symbol σ_i to a latent semantic facet vector, and $\tilde{\mathbf{E}}^{(T)}$ encodes terminal expression embeddings. The aggregation function AGG applies attention-weighted pooling over semantic dimensions, facilitating compositional cultural meaning extraction.

SEA further represents the evolving cultural space as a metric topological manifold:

$$\mathcal{T}_{\text{cult}} = (\mathbb{C}, d_{\text{SEA}}) \tag{35}$$

In this formulation, $\mathbb{C}$ denotes the set of cultural instantiations, and d_{SEA} is a composite metric that includes both symbolic and expression-space divergence:

$$d_{\text{SEA}}(\gamma_i, \gamma_j) = \mathcal{D}_s(\mathbf{y}_i, \mathbf{y}_j) + \lambda_d \cdot \left\|\tilde{\mathbf{E}}_i^{(T)} - \tilde{\mathbf{E}}_j^{(T)}\right\|^2 \tag{36}$$

The symbolic distance $\mathcal{D}_s$ may include tree edit distances between ritual scripts, semantic graph divergence, or attention-weighted lattice mismatches. This metricized space $\mathcal{T}_{\text{cult}}$ enables clustering, morphing, and interpolation between cultural motifs, supporting tasks like recontextualization, ritual adaptation, and mythopoetic transformation.

IV. EXPERIMENTAL SETUP

A. DATASET

We utilize four diverse and culturally significant datasets to evaluate the effectiveness of our approach on intangible cultural heritage (ICH) tasks. The ICH Dataset [19] is a large-scale, multimodal corpus composed of high-quality videos, images, and associated textual descriptions covering more than 100 ICH categories such as traditional crafts, oral expressions, and ritual practices. This dataset provides fine-grained labels and rich contextual metadata, enabling both classification and retrieval tasks. FolkDance DB Dataset [20] focuses on traditional folk dances from different regions, containing annotated video sequences that describe various postures, temporal motion patterns, and regional styles. It supports pose estimation, dance classification, and motion transfer experiments. Chinese Shadow Puppetry Dataset [21] captures performances of Chinese shadow puppetry art with video, audio, and textual narrative data. It includes temporal segmentation, character motion, and audio-visual alignment annotations, making it suitable for cross-modal learning and heritage preservation. UNESCO ICH Video Archive Dataset [22] is compiled from official UNESCO archives, providing authentic video records and documentation of globally recognized intangible cultural heritages. It includes multilingual subtitles, expert commentaries, and detailed cultural metadata. This dataset is essential for evaluating cross-cultural generalization, temporal grounding, and understanding of real-world ICH practices across different geographies and traditions.

B. EXPERIMENTAL DETAILS

The experiments were performed on a workstation featuring NVIDIA A100 GPUs with 40GB of memory, 256GB of RAM, and Intel Xeon Gold 6326 processors. The operating system is Ubuntu 22.04, and the deep learning framework used is PyTorch 2.0.1. We utilize CUDA 12.1 and cuDNN 8.9 for GPU acceleration. All models are implemented using mixed-precision training (AMP) to reduce memory footprint and accelerate convergence. For training, we adopt the AdamW optimizer with a base learning rate of 3e-4, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and weight decay set to 0.01. A linear warmup of 10% total steps is used, followed by cosine decay. For fair comparisons across datasets, batch sizes are dynamically adjusted depending on the input resolution: 32 for 224×224, and 16 for 384×384 input size. During finetuning, learning rates are decayed to 1e-5. Each experiment is run with five random seeds to ensure statistical robustness, and mean with standard deviation is reported. The input modalities vary per dataset. For video inputs, we sample 8 to 16 frames uniformly from each clip, and use temporal augmentations such as frame jittering and temporal cropping. For image inputs, we apply random resized crop (scale=[0.8, 1.0]), horizontal flipping, and color jittering. Textual inputs are tokenized using SentencePiece with a vocabulary size of 32,000 tokens. For audio-related tasks (in datasets such as Chinese Shadow Puppetry), we extract 128-bin log Mel spectrograms using a 25ms window and 10ms stride. Cross-modal experiments use

either late fusion (via transformer cross-attention) or early fusion (via modality token concatenation), depending on task configuration. Our backbone encoder is a ViT-B/16 model pretrained on ImageNet-21K, followed by a task-specific projection head. For video, we use a TimeSformer variant that models spatial and temporal attention separately. Text inputs are encoded using a RoBERTa-base model, and for audio we employ a CNN14 encoder pretrained on AudioSet. During multi-modal pretraining, we apply contrastive loss with temperature scaling, masked modeling loss, and cross-modal matching loss jointly optimized with total loss coefficients set to 1.0, 0.5, and 0.5 respectively. Evaluation metrics include top-1 and top-5 accuracy for classification, mean average precision (mAP) for retrieval, BLEU and CIDEr scores for generation tasks, and F1/IoU for segmentation. For qualitative analysis, t-SNE and Grad-CAM visualizations are provided to interpret cross-modal alignment and attention distributions. For cross-dataset generalization tests, we train on one dataset and test on another under zero-shot settings, without any finetuning. Ablation studies are conducted by removing individual modules such as cross-attention layers, modality-specific encoders, or pretraining objectives to measure their contributions independently.

C. COMPARISON WITH SOTA METHODS

We assess the performance and adaptability of the proposed HoloCultura approach on a range of datasets related to intangible cultural heritage, we compare its performance with a series of state-of-the-art (SOTA) models including CLIP [23], I3D [24], Timesformer [25], SlowFast [26], VideoMAE [27], and ViViT [28]. We report Accuracy, Recall, F1 Score, and AUC on four benchmark datasets as shown in Table 1 and Table 2. HoloCultura achieves the best performance across all metrics and datasets, highlighting its robustness and adaptability to culturally rich, multi-modal content. On the ICH Dataset, HoloCultura outperforms the next-best method, VideoMAE, by a margin of +3.73% in Accuracy and +3.32% in F1 Score. Similarly, on the FolkDance DB Dataset, HoloCultura shows a gain of +5.22% in Accuracy and +4.93% in AUC over CLIP, the strongest baseline. These consistent improvements confirm HoloCultura's advantage in processing fine-grained, motion-heavy video content and cultural-specific contextual cues.

A deeper analysis reveals that traditional vision-language models such as CLIP struggle with nuanced cultural semantics that require modeling complex motion dynamics, object interactions, and temporal coherence. Although CLIP benefits from large-scale pretraining, it lacks specialized temporal reasoning which is critical for tasks such as dance classification or puppet performance interpretation. Video-centric models like I3D and SlowFast exhibit slightly better performance than CLIP in the FolkDance DB and Shadow Puppetry datasets due to their explicit temporal encoding. However, these architectures rely heavily on dense 3D convolutions, leading to higher computation costs and limited generalization in cross-modal tasks. Transformer-based methods such as Timesformer and ViViT provide relatively better scalability and sequence modeling, but their lack of strong modality alignment leads to suboptimal F1 Scores. In contrast, HoloCultura leverages hierarchical cross-modal fusion and temporal-guided attention mechanisms, allowing it to capture spatiotemporal context and cross-modal alignment more effectively. This is evident in its superior Recall and AUC on the UNESCO ICH Video Archive Dataset, where high inter-class similarity and low inter-instance variance pose significant challenges to most baselines. In particular, the gains on the Chinese Shadow Puppetry Dataset are substantial: HoloCultura exceeds VideoMAE by +4.65% in Accuracy and +3.41% in F1 Score. This is largely attributed to HoloCultura's architecture that integrates both low-level motion cues and high-level semantic embeddings, allowing better modeling of hand gesture sequences, stage-object interaction, and audio synchronization which are essential to puppetry analysis. HoloCultura includes a modality-aware alignment strategy that resolves ambiguity when visual content is weakly paired with abstract narratives or traditional vocalizations—issues where methods like ViViT and I3D falter. On the UNESCO ICH dataset, HoloCultura again dominates with +5.29% Accuracy over Timesformer and +6.19% higher AUC over CLIP. These results demonstrate HoloCultura's strong generalization capabilities across cultural boundaries and data distributions.

D. ABLATION STUDY

To further understand the contribution of each core component within our HoloCultura framework, we conduct a comprehensive ablation study by progressively removing major modules. We denote three critical sub-modules as follows: Temporal-Guided Cross-Attention, Modality-Aware Alignment Layer, and Contrastive Pretraining Objective. We evaluate the impact of removing each on all four benchmark datasets. The corresponding quantitative results are reported in Table 3 and Table 4. From the full model (Ours) to each ablated variant (w./o. Symbolic Context Encoding/Hierarchical Generative Semantics/Context-Aware Alignment), we observe consistent performance degradation across all metrics and datasets, which confirms the necessity and complementary benefits of each design.

The removal of Symbolic Context Encoding (w./o. Symbolic Context Encoding), which excludes the temporal-guided cross-attention mechanism, results in the most severe performance drop, particularly on the FolkDance DB and Chinese Shadow Puppetry datasets. This observation is expected, as both datasets are highly motion-centric and rely heavily on precise temporal modeling of body postures or puppet articulation. For example, in FolkDance DB, Accuracy drops from 93.35% to 90.01% and AUC decreases from 94.89% to 90.66%. This underlines the importance of temporal attention in capturing long-term motion dependencies. Similarly, the Chinese Shadow Puppetry Dataset

TABLE 1. Benchmarking our method against SOTA models on the ICH and FolkDance DB collections.

Model	ICH Dataset				FolkDance DB Dataset			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
CLIP [23]	88.62±0.03	86.14±0.02	85.93±0.02	90.27±0.02	87.49±0.02	85.06±0.02	83.77±0.03	88.10±0.03
I3D [24]	85.11±0.02	82.78±0.01	84.66±0.02	87.94±0.03	86.95±0.03	84.72±0.03	82.66±0.02	86.48±0.02
Timesformer [25]	87.32±0.02	84.50±0.03	86.22±0.02	89.31±0.03	88.13±0.02	83.91±0.02	84.25±0.02	87.89±0.03
SlowFast [26]	86.57±0.03	83.34±0.02	85.19±002	88.75±0.02	85.23±0.03	82.15±0.01	81.34±0.02	85.91±0.03
VideoMAE [27]	89.14±0.03	87.62±0.02	86.83±0.02	90.15±0.02	87.62±0.02	84.94±0.02	85.73±0.02	89.02±0.02
ViViT [28]	88.05±0.02	85.77±0.03	84.55±0.02	89.63±0.03	86.48±0.02	83.38±0.02	82.49±0.03	87.44±0.02
Ours (HoloCultura)	**92.87±0.02**	**90.94±0.02**	**89.78±0.02**	**94.22±0.02**	**93.35±0.02**	**91.50±0.02**	**90.66±0.02**	**94.89±0.02**

TABLE 2. Evaluation of the proposed method versus state-of-the-art techniques on the Chinese shadow puppetry and UNESCO ICH video collections.

Model	Chinese Shadow Puppetry Dataset				UNESCO ICH Video Archive Dataset			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
CLIP [23]	85.42±0.03	82.67±0.02	83.15±0.02	86.81±0.03	84.17±0.02	81.95±0.03	80.83±0.03	85.39±0.02
I3D [24]	83.18±0.02	80.49±0.02	81.70±0.02	85.20±0.02	82.61±0.02	80.03±0.02	78.42±0.03	83.91±0.03
Timesformer [25]	84.79±0.02	83.15±0.03	82.04±0.02	86.37±0.02	85.45±0.03	82.11±0.02	83.63±0.02	86.44±0.02
SlowFast [26]	82.93±0.03	81.44±0.02	79.87±0.02	84.76±0.03	83.03±0.02	80.14±0.02	79.63±0.03	84.10±0.03
VideoMAE [27]	86.08±0.02	83.94±0.02	84.25±0.02	87.15±0.02	86.27±0.03	84.55±0.02	84.00±0.02	87.71±0.02
ViViT [28]	84.12±0.02	81.08±0.03	82.94±0.02	86.05±0.02	84.74±0.02	82.23±0.02	81.34±0.02	86.02±0.02
Ours (HoloCultura)	**90.73±0.02**	**88.49±0.02**	**87.66±0.02**	**91.42±0.02**	**91.56±0.02**	**89.11±0.02**	**88.37±0.02**	**92.23±0.02**

exhibits a drop in F1 Score from 87.66% to 84.75%, suggesting that the spatiotemporal dynamics encoded via cross-attention are vital for recognizing cultural gestures and interactions on stage. On the other hand, removing Hierarchical Generative Semantics (w./o. Hierarchical Generative Semantics), the modality-aware alignment layer, also leads to significant performance losses, especially in the UNESCO ICH Video Archive Dataset. In this dataset, where multi-lingual subtitles, heterogeneous narration styles, and loosely paired visuals are present, the alignment layer plays a crucial role in bridging modality gaps. Its removal causes a decrease in AUC from 92.23% to 90.83% and F1 Score from 88.37% to 86.39%, confirming that HoloCultura's fine-grained alignment strategy is essential for dealing with noisy, weakly supervised cross-modal correspondence. The ICH Dataset shows Accuracy reduction of −1.82% under this configuration, indicating that the module also helps in grounding narrative semantics into visual features. We examine the impact of removing the contrastive pretraining objective (w./o. Context-Aware Alignment). This results in a more moderate yet consistent drop across datasets. Since contrastive learning enhances feature separability and modality-agnostic representation space, its absence slightly reduces generalization capability. For instance, in the ICH Dataset, F1 Score drops from 89.78% to 87.18%, while in the Shadow Puppetry Dataset, Accuracy reduces from 90.73% to 88.42%. Though the drop is less severe than the exclusion of Symbolic Context Encoding or Hierarchical Generative Semantics, this result demonstrates that contrastive objectives are effective in optimizing shared representation spaces and reinforcing modality coupling.

Table 5 reports the per-run performance results of the proposed HoloCultura model on the ICH Dataset across five independent runs with different random seeds. The key evaluation metrics include Accuracy, Recall, F1 Score, and AUC. The results demonstrate a high level of consistency across different runs. The Accuracy ranges from 92.75% to 92.95%, with a mean of 92.87% and a standard deviation of 0.08. Similarly, Recall varies between 90.80% and 91.10%, F1 Score fluctuates from 89.60% to 89.95%, and AUC remains tightly concentrated between 94.10% and 94.30%. The low standard deviation across all evaluation metrics indicates that the proposed method exhibits high stability and robustness under different initialization conditions. This consistency reflects the effectiveness and generalizability of the HoloCultura framework in modeling and classifying ICH data. Furthermore, the minimal performance fluctuation across runs suggests that the training process is resilient to random seed variations, ensuring reproducibility of the reported results.

TABLE 3. Evaluation of individual module contributions via ablation studies on ICH and FolkDance DB.

Model	ICH Dataset				FolkDance DB Dataset			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
w./o. Symbolic Context Encoding	89.32±0.03	87.14±0.02	86.05±0.02	90.48±0.03	90.01±0.02	87.30±0.02	85.99±0.02	90.66±0.02
w./o. Hierarchical Generative Semantics	91.05±0.02	88.92±0.02	88.03±0.03	92.01±0.02	91.47±0.03	89.02±0.02	87.14±0.02	91.70±0.03
w./o. Context-Aware Alignment	90.41±0.02	89.27±0.03	87.18±0.02	91.14±0.02	91.06±0.02	88.11±0.03	87.55±0.02	91.89±0.02
Ours	**92.87±0.02**	**90.94±0.02**	**89.78±0.02**	**94.22±0.02**	**93.35±0.02**	**91.50±0.02**	**90.66±0.02**	**94.89±0.02**

TABLE 4. Module-wise ablation results on Chinese shadow puppetry and UNESCO ICH video datasets.

Model	Chinese Shadow Puppetry Dataset				UNESCO ICH Video Archive Dataset			
	Accuracy	Recall	F1 Score	AUC	Accuracy	Recall	F1 Score	AUC
w./o. Symbolic Context Encoding	87.61±0.02	85.38±0.02	84.75±0.03	88.19±0.03	88.05±0.02	85.61±0.02	85.03±0.02	89.35±0.02
w./o. Hierarchical Generative Semantics	89.14±0.03	86.70±0.02	86.09±0.02	89.71±0.02	89.77±0.03	87.44±0.02	86.39±0.02	90.83±0.02
w./o. Context-Aware Alignment	88.42±0.02	86.90±0.02	85.11±0.02	88.91±0.03	90.25±0.02	87.82±0.03	87.05±0.02	91.32±0.03
Ours	**90.73±0.02**	**88.49±0.02**	**87.66±0.02**	**91.42±0.02**	**91.56±0.02**	**89.11±0.02**	**88.37±0.02**	**92.23±0.02**

TABLE 5. Per-run performance results on the ICH dataset (5 independent runs).

Run Index	Accuracy (%)	Recall (%)	F1 Score (%)	AUC (%)
Run 1	92.75	90.80	89.60	94.10
Run 2	92.90	91.05	89.85	94.30
Run 3	92.95	91.10	89.80	94.25
Run 4	92.85	90.90	89.70	94.20
Run 5	92.95	91.00	89.95	94.25
Mean	**92.87**	**90.94**	**89.78**	**94.22**
Std	**0.08**	**0.11**	**0.13**	**0.07**

V. DISCUSSION

While the proposed HoloCultura framework represents a significant technical advance in modeling, preserving, and interactively transmitting ICH, we recognize important epistemological and ethical limitations inherent in computational approaches to cultural representation. Despite incorporating mechanisms like diachronic memory and context-sensitive alignment, our model still operates within structured, quantifiable frameworks that assume a degree of symbolic fixity and semantic computability. As critical scholars such as Assmann and Trouillot argue, cultural meaning is not merely data to be abstracted but is continuously contested, negotiated, and socially lived. We acknowledge that the abstraction and categorization required for machine learning may risk flattening complex cultural dynamics, reifying dominant narratives, or ignoring marginalized voices. To address this, we advocate for future research directions that integrate reflexive design principles, community-driven validation protocols, and participatory knowledge governance mechanisms. These approaches will help ensure that computational representations remain aware of their own limitations and are subjected to ongoing human critique and re-interpretation. This reflexivity is essential for advancing interdisciplinary knowledge production at the intersection of AI, cultural studies, and heritage conservation.

VI. CONCLUSION AND FUTURE WORK

In this work, we address the pressing challenge of preserving Intangible Cultural Heritage (ICH), which is inherently fluid, performative, and often passed down orally. Traditional preservation methods have struggled to capture these ephemeral and context-dependent characteristics. To overcome these limitations, we developed a novel computational framework that blends advanced neural representations with structured symbolic logic and contextual reasoning. Central to our method is a neural-symbolic architecture that captures the multimodal and evolving nature of cultural expressions. We also propose a culturally informed reasoning strategy that employs a self-supervised semiotic alignment module, enabling our system to align and adapt to both canonical and emergent forms of ICH through interactive learning and contextual engagement. Experiments demonstrate that our

approach provides a dynamic and robust means for digitally presenting and engaging with ICH, ensuring its long-term accessibility and vitality through interactive AI-driven tools.

To further align with decolonial heritage perspectives and ethical AI principles, the integration of the HITL feedback module marks a step towards repatriating interpretive authority over cultural representations. By embedding community-centered co-authorship into the modeling pipeline, our framework promotes epistemic justice and empowers practitioners to critically engage with, refine, or reject AI-generated outputs. This participatory approach fosters dialogic knowledge construction, reduces the risk of symbolic essentialism, and ensures that the preservation and digital modeling of ICH remain inclusive, context-sensitive, and accountable to the communities they represent.

To mitigate the risk of overfitting to hegemonic or overly institutionalized cultural prototypes, future iterations of the SEA framework will incorporate symbolic entropy analysis and divergence-tolerance mechanisms. By computing entropy across symbolic alignment layers, the model will gain the capacity to quantify and respond to distributional variability within cultural observations. High-entropy signals, often characteristic of diasporic, hybrid, or emergent cultural forms, will trigger adaptive thresholding in the alignment scoring functions. We plan to introduce context-sensitive innovation metrics, designed to assess symbolic recombination rates, stylistic deviation, and the emergence of novel cultural motifs. These metrics will serve as positive indicators in contexts where cultural creativity and heterogeneity are valued. Moreover, divergence-aware loss regularization will be explored to prevent excessive alignment convergence, thus maintaining sensitivity to legitimate forms of cultural variation. These enhancements will ensure that the SEA framework remains both precise and inclusive, capable of modeling both continuity and innovation in intangible cultural heritage.

We also recognize important ethical and epistemological concerns associated with computationally modeling complex cultural semiotics. Despite the adaptive and context-sensitive mechanisms embedded within SEA, our framework inevitably operates within formalized representational boundaries, which may risk oversimplifying culturally nuanced meanings. Particularly for marginalized or underrepresented communities, such modeling may inadvertently reinforce dominant narratives if not carefully designed. Future work will prioritize integrating participatory knowledge co-construction mechanisms, allowing cultural practitioners and community members to directly influence model training, interpretation layers, and alignment criteria. We plan to explore probabilistic uncertainty quantification and multi-hypothesis reasoning strategies to represent multiple coexisting interpretations more faithfully. By embedding human-in-the-loop validation processes and engaging with diverse cultural stakeholders, we aim to mitigate epistemological limitations and foster more inclusive, dialogic, and ethically grounded digital preservation of intangible cultural heritage.

Future developments of the SEA framework will address the ethical tensions arising from the modeling and dissemination of culturally sensitive or situationally restricted intangible cultural heritage (ICH) elements. Recognizing that certain songs, rites, or symbolic practices are intended for specific audiences or ritual contexts, we plan to implement a Community-Defined Access Control Layer. This module will allow community stakeholders to specify access permissions, visibility thresholds, and symbolic data expiration timelines at both data ingestion and output retrieval stages. Role-based authentication protocols will restrict access to sensitive symbolic layers, ensuring that only authorized users with culturally sanctioned roles can view or interact with specific heritage elements. We will explore dynamic privacy-preserving modeling techniques, such as context-aware content gating and encryption-based output filtering, to prevent unauthorized analysis or dissemination of culturally restricted data. These enhancements aim to ensure that our system supports ethical data stewardship while respecting the epistemic sovereignty and situational privacy norms upheld by heritage communities.

Another key area for future development involves addressing the data sparsity, imbalance, and linguistic vulnerability inherent in many ICH contexts. Recognizing that numerous cultural practices remain undocumented, endangered, or exist in low-resource languages, we propose integrating a few-shot and zero-shot generalization module within HoloCultura. This module will leverage pre-trained cross-modal transformers and symbolic meta-reasoning mechanisms to improve performance in scenarios with minimal labeled data. We plan to implement a symbolic bootstrapping pipeline that utilizes community-crowdsourced metadata and annotation inputs to enrich the model's symbolic grounding space. This participatory data augmentation approach will enable practitioners to co-create contextual labels, descriptive tags, and folk taxonomies that help capture underrepresented cultural forms. To mitigate risks of reinforcing existing data biases, we will adopt adversarial de-biasing techniques, domain adaptation strategies, and entropy-based class balancing methods during model training. Moreover, a data auditing dashboard will be developed to enable stakeholders to inspect training distributions, monitor representational equity, and guide corrective data curation. These enhancements aim to promote inclusivity, fairness, and robustness in ICH modeling, particularly for marginalized and vulnerable cultural traditions.

A significant area for future research involves extending the diachronic memory module to account for the socially selective and politically negotiated dimensions of cultural memory, as emphasized in works by Assmann and Trouillot. To capture phenomena such as selective forgetting, symbolic marginalization, and revisionist reinvention, we plan to develop a selective memory attenuation mechanism that

allows certain latent representations within the diachronic memory bank to decay, be suppressed, or be reweighted based on community-defined salience scores, social feedback signals, or external historical indicators. A symbolic marginalization layer will enable dynamic attenuation of low-salience or politically de-emphasized symbolic variants. We also intend to implement context-sensitive memory gating functions, allowing the system to modulate memory retrieval biases in response to shifting socio-cultural contexts. These enhancements will bring our model closer to the nuanced, contested, and dynamic nature of cultural memory in real-world heritage processes.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Liuxun Zhang: conceptualization, software, and writing—review and editing; Zhouluo Wang: formal analysis, investigation, and data curation; Rulan Yang: methodology; Qiang Yi: validation, visualization, supervision, and funding acquisition; and Liuxun Zhang, Zhouluo Wang, Rulan Yang, and Qiang Yi: writing—original draft preparation. All authors have reviewed and approved the final version of the manuscript for publication.

ACKNOWLEDGMENT

(Liuxun Zhang and Zhouluo Wang contributed equally to this work.)

REFERENCES

[1] K. Luxem, J. J. Sun, S. P. Bradley, K. Krishnan, E. Yttri, J. Zimmermann, T. D. Pereira, and M. Laubach, "Open-source tools for behavioral video analysis: Setup, methods, and best practices," *eLife*, vol. 12, Mar. 2023, Art. no. e79305.

[2] S. Wan, X. Xu, T. Wang, and Z. Gu, "An intelligent video analysis method for abnormal event detection in intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4487–4495, Jul. 2021.

[3] D. Kitaguchi, N. Takeshita, H. Matsuzaki, T. Igaki, H. Hasegawa, and M. Ito, "Development and validation of a 3-Dimensional convolutional neural network for automatic surgical skill assessment based on spatiotemporal video analysis," *JAMA Netw. Open*, vol. 4, no. 8, Aug. 2021, Art. no. e2120786.

[4] W. Liu, G. Kang, P.-Y. Huang, X. Chang, L. Yu, Y. Qian, J. Liang, L. Gui, J. Wen, P. Chen, and A. G. Hauptmann, "Argus: Efficient activity detection system for extended video analysis," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Mar. 2020, pp. 126–133.

[5] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3156–3165.

[6] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, Jan. 2021.

[7] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the 'video' in video-language understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2907–2917.

[8] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12922–12943, Nov. 2023.

[9] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proc. IEEE*, vol. 109, no. 11, pp. 1838–1863, Nov. 2021.

[10] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 2259–2322, Mar. 2021.

[11] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Trans. Image Process.*, vol. 29, pp. 8680–8695, 2020.

[12] C. Wang, S. Zhang, Y. Chen, Z. Qian, J. Wu, and M. Xiao, "Joint configuration adaptation and bandwidth allocation for edge-based real-time video analytics," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Jul. 2020, pp. 257–266. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9155524/

[13] G. Awad, A. Butt, K. Curtis, J. G. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. L. Diduch, J. Liu, A. Smeaton, Y. Graham, G. J. Jones, W. Kraaij, and G. Quénot, "TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains," *TREC Video Retr. Eval.*, 2021.

[14] M. Noetel, S. Griffith, O. Delaney, T. Sanders, P. Parker, B. del Pozo Cruz, and C. Lonsdale, "Video improves learning in higher education: A systematic review," *Rev. Educ. Res.*, vol. 91, no. 2, pp. 204–236, Apr. 2021.

[15] F. Yuanta, "Pengembangan media video pembelajaran ilmu pengetahuan sosial pada siswa sekolah dasar," *Trapsila: Jurnal Pendidikan Dasar*, vol. 1, no. 2, p. 91, Feb. 2020.

[16] M. Aloraini, M. Sharifzadeh, and D. Schonfeld, "Sequential and patch analyses for object removal video forgery detection and localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 917–930, Mar. 2021.

[17] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Netw. Anal. Mining*, vol. 11, p. 81, Aug. 2021.

[18] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, and J. P. McCrae, "Corpus creation for sentiment analysis in code-mixed Tamil-English text," in *Proc. Workshop Spoken Lang. Technol. Under-Resourced Lang.*, 2020, pp. 1–9.

[19] D. Ma, C. Li, T. Du, L. Qiao, D. Tang, Z. Ma, L. Shi, G. Lu, Q. Meng, Z. Chen, M. Grzegorzek, and H. Sun, "PHE-SICH-CT-IDS: A benchmark CT image dataset for evaluation semantic segmentation, object detection and radiomic feature extraction of perihematomal edema in spontaneous intracerebral hemorrhage," *Comput. Biol. Med.*, vol. 173, May 2024, Art. no. 108342.

[20] Z. Miao, W. Wang, J. Xie, L. Ma, and N. Hu, "Research on original environment folk dance movement evaluation based on spatio-temporal graph convolutional networks," *Signal, Image Video Process.*, vol. 19, no. 3, p. 266, Mar. 2025.

[21] Y. Tong, J. Xu, X. Du, J. Huang, and H. Zhou, "SP-GAN: Cycle-consistent generative adversarial networks for shadow puppet generation," in *Proc. IEEE Int. Conf. Cybern. Intell. Syst. (CIS) IEEE Int. Conf. Robot., Autom. Mechatronics (RAM)*, Aug. 2024, pp. 32–38.

[22] H. Tahvanainen, T. Ylönen, and O. Valo, "Feature comparison for classification of kaustinen fiddle playing style from archived recordings using deep learning," in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2024, pp. 1007–1011.

[23] B. Zhang, P. Zhang, X. Dong, Y. Zang, and J. Wang, "Long-CLIP: Unlocking the long-text capability of CLIP," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2024, pp. 310–325. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-72983-6_18

[24] T. Moodley and D. van der Haar, "I3D-AE-LSTM: Combining action representations using a 2-stream autoencoder for action quality assessment," *Expert Syst. Appl.*, vol. 278, Jun. 2025, Art. no. 127368.

[25] Z. Chen, S. Wang, D. Yan, and Y. Li, "A spatio–temporl deepfake video detection method based on timesformer-CNN," in *Proc. 3rd Int. Conf. Distrib. Comput. Elect. Circuits Electron. (ICDCECE)*, Apr. 2024, pp. 1–6.

[26] M. Munsif, N. Khan, A. Hussain, M. J. Kim, and S. W. Baik, "Darkness-adaptive action recognition: Leveraging efficient tubelet slow-fast network for industrial applications," *IEEE Trans. Ind. Informat.*, vol. 20, no. 12, pp. 13676–13686, Dec. 2024.

[27] J. Moon, S. Heo, J. Won, J. Jang, and S. K. Jung, "State space model based VideoMAE enhancement for efficient video action classification," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC)*, Feb. 2025, pp. 0820–0824.

[28] T. Higashi, R. Ishibashi, and L. Meng, "ViViT fall detection and action recognition," in *Proc. Int. Conf. Adv. Mech. Syst. (ICAMechS)*, Nov. 2024, pp. 291–296.

LIUXUN ZHANG received the B.A. degree in journalism from Beijing Foreign Studies University, Beijing, China, and the M.A. degree in communication from Guangxi Science Technology Normal University, China.

He is currently a Lecturer with the School of Literature and Media, Guangxi Science Technology Normal University. He has published several articles in national journals and participated in projects related to media globalization and multilingual news production. His research interests include international journalism, cross-cultural communication, and media discourse analysis.

Mr. Zhang has participated in multiple academic symposiums and serves as a Reviewer for *Journal of International Communication*.

ZHOULUO WANG received the B.A. degree in journalism from Beijing Sport University, Beijing, China, and the M.A. degree in marketing communications from The University of Melbourne, Melbourne, Australia. She is currently pursuing the Ph.D. degree with the School of Philosophy and Sociology, Jilin University, Changchun, China.

Her academic interests include intangible cultural heritage, sports society, and culture, with a particular focus on utilizing technology to promote the safeguarding of traditional culture. She has published several articles and a monograph on sports intangible cultural heritage.

Ms. Wang has presented at international forums on sports science and contributed to social science foundation projects.

RULAN YANG received the B.A. degree in international journalism from Beijing Foreign Studies University, Beijing, China, where she is currently pursuing the bachelor's degree with the School of International Journalism.

Her academic interests include global communication, media narratives, and transnational news reporting. She has engaged in collaborative research on international media and has contributed to projects focusing on cross-border media practices.

Ms. Yang is actively involved in student-led academic forums and has assisted in organizing university-level communication workshops.

QIANG YI received the Ph.D. degree in communication studies from National Chengchi University, Taiwan.

He is currently a Professor at the School of Literature and Communication, Quanzhou Normal University, Fujian, China, and a Visiting Researcher at the School of Communication, National Chengchi University. He has authored several academic monographs and more than 40 journal articles. His research interests include media culture, political communication, and comparative media systems, with a focus on the role of media in cross-strait relations and East Asian discourse systems.

Prof. Yi was a recipient of multiple provincial-level research grants and serves as a Reviewer for journals such as *Asian Journal of Communication* and *Global Media and China*.