scientific reports

Check for updates

OPEN

Knowledge graph enhanced cross modal generative adversarial network for martial arts motion reconstruction and heritage preservation

Xiaoyu Yue✉ & Lulu Zhang

This paper presents a novel knowledge graph enhanced cross-modal generative adversarial network (KG-CMGAN) for preserving traditional martial arts techniques. We address the challenges of capturing the complex, multidimensional nature of martial arts by integrating structured domain knowledge with advanced deep learning architectures. Our framework establishes an end-to-end solution that bridges visual, textual, and sequential representations to achieve comprehensive motion reconstruction while preserving stylistic authenticity and semantic meaning. The proposed approach includes a comprehensive martial arts knowledge graph that formalizes domain-specific ontology, a knowledge-guided cross-modal alignment mechanism that effectively integrates heterogeneous data sources, and a knowledge-enhanced adversarial learning architecture specifically optimized for martial arts motion reconstruction. Extensive experiments across six traditional Chinese martial arts styles demonstrate significant improvements over state-of-the-art baselines, with 28.4% reduction in joint position error and 91.2% knowledge consistency score. Ablation studies confirm that knowledge graph integration is critical for generating culturally authentic movements. This research contributes a novel methodology for intangible cultural heritage preservation that captures both the physical execution and conceptual foundations of traditional martial arts.

Keywords Knowledge graph, Martial arts, Cross-modal learning, Motion reconstruction, Generative adversarial networks, Cultural heritage preservation

Traditional martial arts, as a significant component of intangible cultural heritage, embody the profound cultural wisdom and physical intelligence accumulated over centuries of human civilization[1]. Chinese martial arts (Wushu), in particular, with their history dating back more than 4000 years, represent not merely a collection of combat techniques but also encompass philosophical principles, health preservation methods, and aesthetic expressions unique to Eastern culture[2]. However, the inheritance of martial arts techniques faces unprecedented challenges in contemporary society, including the aging of technique inheritors, the fragmentation of knowledge transmission systems, and the lack of standardized documentation methodologies that can comprehensively capture the nuanced movements and implicit knowledge embedded in these practices[3].

Conventional approaches to preserving martial arts techniques have predominantly relied on apprenticeship models and textual documentation, which are inherently limited in their capacity to record the dynamic, three-dimensional nature of martial movements[4]. Even with the advent of video recording technology, significant information loss occurs in the translation of complex spatiotemporal movements to two-dimensional representations. Moreover, these methodologies fail to capture the underlying biomechanical principles, energy flow patterns, and tactical considerations that constitute the essence of martial arts expertise[5].

Recent advancements in motion capture and computer vision technologies have enabled more precise recording of human movements. However, these technologies present substantial limitations when applied to martial arts practice. Commercial motion capture systems typically require specialized environments and equipment that restrict natural movement expression, while vision-based approaches struggle with occlusion issues, rapid movements, and the extraction of fine-grained details crucial to martial arts technique mastery[6].

College of Physical Education, Nanjing Tech University, Nanjing 211816, JiangSu, China. ✉email: yuexiaoyu2025@163.com

Furthermore, existing systems operate primarily as documentation tools, lacking the interpretative frameworks necessary to understand the semantic significance and contextual applications of recorded movements[7].

The intersection of artificial intelligence, multimodal learning, and knowledge representation offers promising avenues to address these challenges. Knowledge graphs, as structured representations of domain-specific information, have demonstrated remarkable capability in organizing complex relational data across numerous fields[8]. When integrated with deep learning architectures, particularly generative adversarial networks (GANs), they present unprecedented opportunities for not only preserving but also understanding and reconstructing martial arts techniques in their full complexity[9].

This paper proposes a novel Knowledge Graph-Enhanced Cross-Modal Generative Adversarial Network (KG-CMGAN) framework for martial arts technique inheritance. Our approach establishes an end-to-end solution that bridges visual, textual, and sequential representations to achieve comprehensive motion reconstruction. The framework leverages the complementary strengths of different modalities: visual data captures external movement forms, textual descriptions provide semantic context and technical principles, while sequential representations encode the temporal dynamics of techniques[10].

The significant contributions of this research are threefold:

1. We develop a comprehensive martial arts knowledge graph that formalizes the domain-specific ontology of techniques, principles, applications, and biomechanical features, establishing a structured foundation for intelligent motion analysis.
2. We design a novel cross-modal fusion mechanism that effectively integrates information from heterogeneous data sources (video, text, motion sequence), enabling more robust feature representation and addressing the inherent limitations of single-modality approaches.
3. We implement an adversarial learning architecture specifically optimized for martial arts motion reconstruction, incorporating domain knowledge constraints to ensure both physical plausibility and stylistic authenticity in generated movements.

Beyond its technical innovations, this research holds profound significance for cultural heritage preservation. By creating a digital framework capable of capturing, interpreting, and reconstructing martial arts techniques with unprecedented fidelity, we establish a sustainable approach to safeguarding these invaluable cultural practices for future generations. Additionally, the methodology proposed here has potential applications extending beyond martial arts to other forms of intangible cultural heritage characterized by complex motion patterns and rich contextual knowledge. Figure 1 illustrates the overall architecture of our proposed KG-CMGAN framework, showing the data flow from multimodal inputs through knowledge-guided processing to motion reconstruction, with explicit visualization of where knowledge graph constraints are applied and how different loss functions govern the learning dynamics.

The framework consists of five main components: (1) multimodal encoders: video encoder (3D CNN with spatio-temporal attention), text encoder (domain-adapted Transformer), and motion sequence encoder (spatial-temporal GCN) extract features from their respective modalities; (2) knowledge layer: The martial arts knowledge graph is processed by RGCN to generate entity embeddings Z, which guide the alignment and generation processes; (3) cross-modal fusion: KNOWLEDGE-guided attention mechanisms align features from different modalities into a unified 768-dimensional representation space; (4) generation module: transformer-based generator with knowledge-conditioned attention produces motion sequences, with knowledge

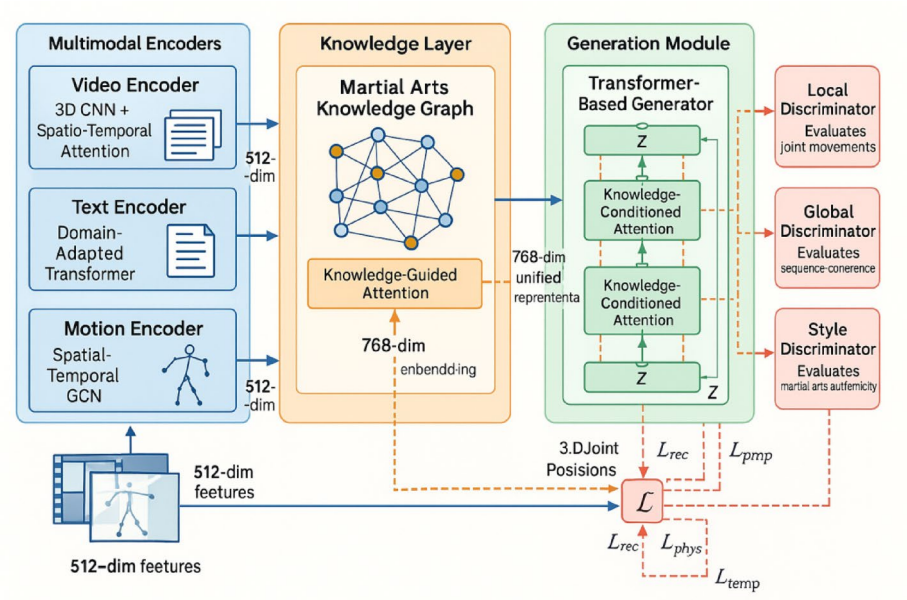

Fig. 1. Overall architecture of the KG-CMGAN framework.

embeddings Z injected at multiple layers; (5) discrimination module: three discriminators (local, global, style) evaluate generated motions at different granularities, each conditioned on knowledge graph information. The five loss terms ($\mathcal{L}_{GAN}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{kg}$, $\mathcal{L}_{phys}$, $\mathcal{L}_{temp}$) are computed at corresponding modules to jointly optimize the framework. Blue arrows indicate data flow, orange dashed lines show knowledge guidance paths, and red dotted lines represent gradient backpropagation.

Related work and theoretical foundation

Current research status of martial arts motion capture and reconstruction technology

Motion capture and reconstruction technologies for martial arts have evolved significantly over the past decades, transitioning from rudimentary video documentation to sophisticated multi-sensor systems and AI-driven approaches. Early efforts primarily relied on traditional videography supplemented by manual annotation, which provided basic archival capabilities but failed to capture the three-dimensional complexity and biomechanical nuances essential to martial arts movements[11]. This methodological limitation significantly hindered the accurate preservation and analytical study of martial arts techniques, particularly for styles characterized by explosive movements and subtle weight distributions.

The emergence of marker-based motion capture systems marked a substantial advancement in the field, enabling precise tracking of joint positions and orientations through retroreflective markers and multiple calibrated cameras[12]. While these systems achieved millimeter-level accuracy under controlled conditions, their application to martial arts documentation revealed significant constraints. The physical presence of markers often restricts the natural execution of techniques, especially in martial arts forms requiring full-body contact or utilizing traditional weapons[13]. Additionally, the controlled laboratory environment necessary for optimal marker tracking fundamentally alters the contextual authenticity of martial arts practice, potentially compromising the ecological validity of captured data.

Markerless vision-based systems emerged as a promising alternative, utilizing advanced computer vision algorithms to estimate human pose from RGB or RGB-D camera inputs without physical attachments to performers[14]. These approaches have demonstrated substantial potential for non-intrusive motion documentation, yet continue to struggle with fundamental challenges when applied to martial arts contexts. Current vision-based systems exhibit significant performance degradation when processing high-velocity movements characteristic of martial arts techniques, frequently resulting in motion blur and tracking failures during critical technique transitions[15]. Moreover, the occlusion problems inherent in complex martial arts movements, particularly in techniques involving close-quarters interactions or intricate limb configurations, remain largely unresolved in contemporary vision-based frameworks[16].

Inertial measurement unit (IMU) based systems represent another technological approach, utilizing body-worn sensors containing accelerometers, gyroscopes, and magnetometers to track segment orientations and reconstruct full-body movements[17]. While these systems offer greater mobility and independence from optical constraints, they introduce cumulative drift errors over extended recording periods and struggle with absolute position estimation, particularly problematic for techniques requiring precise spatial positioning and directional awareness[18]. The integration challenges between multiple IMUs further complicate the capture of whole-body coordination patterns essential to martial arts performance.

Recent deep learning approaches have attempted to address these limitations through end-to-end motion reconstruction frameworks[19]. Convolutional neural networks and recurrent architectures have demonstrated improved robustness in pose estimation from monocular video, while transformer-based models have enhanced temporal coherence in movement sequence prediction[20]. Despite these advancements, current AI-driven approaches remain predominantly trained on general human movement datasets, lacking the specialized knowledge required to accurately interpret the unique biomechanical patterns and stylistic nuances distinctive to different martial arts traditions[21]. For instance, while general pose estimation models achieve high accuracy on benchmark datasets like Human3.6 M, they frequently fail to capture the subtle weight distribution shifts and internal energy flow patterns essential to Taijiquan techniques, or the explosive whole-body coordination characteristic of Xingyiquan movements. This performance gap highlights the necessity for domain-specific knowledge integration rather than relying solely on data-driven learning approaches.

A fundamental limitation permeating all current technological approaches lies in their predominantly biomechanical focus, which neglects the rich semantic and cultural dimensions embedded within martial arts techniques[22]. Existing systems effectively capture "what" movements occur but fail to encode "why" these movements are executed—the tactical considerations, philosophical principles, and cultural contexts that give meaning to physical forms. This semantic gap severely restricts the educational and preservation value of current documentation technologies, particularly for traditional martial arts systems where physical techniques are inseparable from their conceptual foundations.

The integration of multimodal data sources and knowledge-driven approaches represents a promising direction for addressing these limitations, potentially enabling more comprehensive documentation that captures both physical execution and conceptual understanding of martial arts techniques. However, substantial methodological innovations are required to develop frameworks capable of bridging the persistent gap between quantitative movement data and qualitative martial arts knowledge.

Research progress in cross-modal generative adversarial networks

Cross-modal learning represents a transformative paradigm in artificial intelligence, enabling systems to process, understand, and generate content across different representational domains such as images, text, and time-series data. At the core of cross-modal learning lies the challenge of establishing coherent mappings between distinct representational spaces while preserving semantic consistency and contextual relevance[23]. This capability is

particularly critical for domains like martial arts, where physical movements exist concurrently with verbal instructions, conceptual principles, and cultural contexts.

Generative Adversarial Networks (GANs), first introduced by Goodfellow et al., have revolutionized generative modeling through their adversarial optimization framework[24]. The standard GAN formulation establishes a minimax game between a generator network G and a discriminator network D, represented by the objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$$

Where $p_{data}(x)$ denotes the distribution of real data, and $p_z(z)$ represents the prior distribution of latent variables. This adversarial dynamic drives the generator to produce increasingly realistic outputs while simultaneously training the discriminator to become more discerning in distinguishing genuine from synthetic samples.

The integration of cross-modal capabilities into the GAN framework has yielded significant advancements in conditional generation tasks. Cross-modal GANs typically incorporate conditioning variables from one modality to guide the generation process in another modality[25]. Recent advances in spatio-temporal adversarial learning have demonstrated remarkable effectiveness in handling high-velocity movements and complex temporal dynamics. Zhang et al.[26] proposed adversarial spatio-temporal learning for video deblurring, utilizing 3D convolutions to jointly capture spatial and temporal information across neighboring frames, which directly addresses challenges similar to those encountered in martial arts motion capture. Enhanced spatio-temporal interaction networks[27] have further improved processing efficiency while maintaining superior performance through deep residual architectures combined with temporal correlation modeling. Comprehensive benchmarking studies on multi-cause blur datasets[28] provide standardized evaluation frameworks that inform our approach to handling motion blur in rapid martial arts movements. These advances in spatio-temporal adversarial learning offer important methodological references for our knowledge-guided motion reconstruction framework, particularly in preserving temporal coherence and spatial details during fast movements.

The conditional GAN objective extends the standard formulation to include a conditioning variable c:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x|c)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|c)))]$$

Text-to-image synthesis represents one of the most extensively researched applications of cross-modal GANs, with architectures like AttnGAN and DALL-E demonstrating remarkable capabilities in generating photorealistic images from textual descriptions[29]. These systems typically employ multi-stage refinement processes, where initial coarse visual representations are progressively enhanced through attention mechanisms that establish fine-grained correspondences between textual features and spatial regions[30]. Despite these advances, current text-to-image systems struggle with generating physically coherent human poses and actions, particularly for specialized movement vocabularies like those found in martial arts.

Conversely, vision-to-text translation has seen significant progress through the development of architectures that bridge convolutional visual encoders with autoregressive text decoders[31]. For high-velocity movements and long-range temporal reasoning crucial to martial arts analysis, Transformer architectures as flow learners[32] have demonstrated superior capabilities in capturing distant temporal dependencies in fast action sequences. In video captioning, global-local representation granularity approaches[33] achieve finer cross-modal alignment through hierarchical attention mechanisms, which aligns well with the multi-level semantic requirements of martial arts technique descriptions. Text-as-stochastic-embedding methods for text-video retrieval[34] provide more robust cross-modal matching strategies, helping to address the polysemy and context-dependence inherent in martial arts terminology. These advances in cross-modal alignment and temporal modeling provide important methodological foundations for our knowledge-guided alignment mechanisms.

Recent transformer-based approaches have further enhanced this capability by introducing cross-attention mechanisms that enable more nuanced mappings between visual regions and textual elements. However, these systems typically generate general descriptive content and lack the specialized vocabulary and conceptual frameworks necessary for articulating the technical nuances and theoretical principles underlying martial arts movements.

Sequence modeling within cross-modal frameworks presents unique challenges, particularly for human motion data characterized by high dimensionality, complex temporal dependencies, and physical constraints[35]. Recurrent architectures augmented with adversarial objectives have demonstrated promising results in motion prediction and completion tasks, while more recent transformer-based approaches have improved long-range temporal modeling capabilities[36]. The sequential cross-modal translation objective can be formulated as:

$$\mathcal{L}_{seq} = \mathbb{E}_{x\sim p_{data}(x), y\sim p_{data}(y|x)}[-\log p_G(y|x)] + \lambda_{adv}\mathcal{L}_{adv}(G,D)$$

Where x represents the source sequence, y denotes the target sequence, and λ_{adv} controls the contribution of the adversarial loss component.

In parallel with GAN-based approaches, diffusion models have emerged as powerful alternatives for human motion generation. Motion Diffusion Model (MDM)[37] employs iterative denoising processes to generate high-quality motion sequences, achieving breakthrough results in general action synthesis tasks. VQ-VAE combined with GPT architectures, exemplified by T2M-GPT[38], discretize motion into token sequences for flexible text-driven generation. Fine-grained text-to-motion diffusion models such as Fg-T2M[39] further enhance understanding of complex textual descriptions through hierarchical attention mechanisms. While these

diffusion-based and tokenized transformer methods demonstrate strong capabilities on general human motion datasets like HumanML3D, they primarily rely on large-scale data-driven learning without explicit modeling of domain-specific knowledge. For example, when generating a Baguazhang circular walking sequence, these models may produce visually plausible movements but fail to maintain the continuous circular footwork pattern and coordinated torso rotation that define the style's essence. Similarly, in Taijiquan applications, they often generate movements with correct joint angles but violate the fundamental principle of substantial-insubstantial weight distribution alternation, resulting in technically flawed executions that appear correct to untrained observers but lack martial authenticity. In contrast, our KG-CMGAN framework addresses this limitation by deeply integrating structured martial arts knowledge with the generative process, making it particularly suitable for specialized domains where cultural semantics and biomechanical principles must be preserved. Despite these advancements, current cross-modal GAN architectures and diffusion models face substantial challenges when applied to specialized domains like martial arts motion modeling.

Semantic alignment remains a persistent issue, particularly when translating between modalities with fundamentally different representational structures and varying levels of abstraction[40]. This challenge is exacerbated in martial arts contexts, where physical movements embody abstract concepts like "energy flow" or "root strength" that have no direct visual manifestations but are crucial to technique execution.

Furthermore, existing frameworks struggle with preserving long-range dependencies and structural coherence in high-dimensional temporal data[41]. Martial arts sequences often exhibit complex hierarchical structures, where individual movements gain meaning through their relationship to broader tactical sequences and strategic principles. Current models typically fail to capture these multi-level dependencies, resulting in generated motions that may appear locally plausible but lack global coherence and functional validity.

The integration of domain-specific knowledge representations with cross-modal GAN architectures presents a promising direction for addressing these limitations. By formalizing martial arts concepts, principles, and relationships within structured knowledge frameworks, cross-modal translation processes can be guided by domain-specific constraints and semantic relationships rather than relying solely on statistical patterns learned from limited training data.

Application of knowledge graphs in cultural heritage protection

Knowledge graphs have emerged as a powerful paradigm for structuring, preserving, and reasoning with complex domain knowledge in cultural heritage contexts. These semantic network structures, comprising entities, relationships, and attributes, enable the formalization of implicit cultural knowledge through explicitly defined ontological frameworks[42]. By representing cultural concepts as interconnected knowledge structures rather than isolated data points, knowledge graphs facilitate more nuanced preservation approaches that capture not only cultural artifacts but also their contextual significance, historical evolution, and interconnections with broader cultural systems.

In the domain of intangible cultural heritage preservation, knowledge graphs have demonstrated significant utility in structuring heterogeneous information across multiple representational modalities. Recent projects have successfully applied knowledge graph methodologies to document traditional craftsmanship techniques, folk music traditions, and ritualistic practices, creating comprehensive digital repositories that interconnect procedural knowledge, cultural context, and performance variations[43]. These semantic frameworks enable more sophisticated query capabilities and inferential reasoning compared to conventional documentation approaches, supporting both preservation objectives and educational applications through enhanced knowledge accessibility and interpretability[44].

The application of knowledge graph technologies to martial arts preservation presents unique challenges due to the domain's complex integration of physical techniques, philosophical principles, and cultural contexts. Recent efforts have begun exploring structured knowledge representations in the martial arts domain, demonstrating the community's recognition of this approach's value. For instance, knowledge graphs of Chinese kung fu masters document lineage relationships, master-disciple connections, and stylistic evolution, providing semanticized resources for martial arts historical and cultural research[45]. While such systems primarily focus on biographical and Inheritance relationships, and our work concentrates on technique-level biomechanical constraints, these parallel efforts collectively validate the broader need for structured martial arts knowledge representation. Similar knowledge engineering projects have constructed semantic networks from different dimensions, enriching our understanding of how computational methods can preserve martial arts heritage.

Knowledge graphs provide particularly valuable support for semantic understanding in multimodal martial arts documentation frameworks. By establishing explicit relationships between terminology, visual demonstrations, and movement sequences, knowledge graphs create structured bridges between different representational modalities, addressing the semantic gap that frequently undermines conventional documentation approaches[46]. This semantic scaffolding enables more coherent interpretations of martial arts content across modalities, facilitating both analysis tasks (extracting semantic meaning from visual demonstrations) and synthesis operations (generating visually accurate movements from textual descriptions).

In cross-modal alignment contexts, knowledge graphs function as intermediary representational structures that establish common semantic reference points across otherwise heterogeneous data sources[47]. This capability proves especially valuable for martial arts documentation, where terminology often exhibits high degrees of cultural specificity and conceptual abstraction. By anchoring diverse representational formats to shared conceptual entities within a knowledge graph, systems can establish more reliable correspondences between technical terms, visual demonstrations, and biomechanical patterns[48].

Furthermore, knowledge graphs provide structured prior knowledge that can significantly enhance the biomechanical and functional plausibility of reconstructed martial arts movements. Considering the practical deployment needs of cultural institutions with limited computational resources, automatic multi-step distillation

methods for large-scale vision models[49] offer feasible pathways for deploying knowledge-enhanced models in resource-constrained environments. While our current work focuses on achieving optimal reconstruction quality, these compute-aware training strategies will be explored in future research to lower deployment barriers for martial arts teaching and exhibition scenarios in museums and cultural centers.

By encoding physical constraints, anatomical relationships, and functional principles within the knowledge structure, reconstruction algorithms can incorporate domain-specific constraints that guide generative processes toward physically valid and martially authentic movement patterns[50]. This approach represents a significant advancement over purely data-driven methods that lack explicit martial arts domain knowledge, particularly for techniques with limited training examples or complex underlying principles.

The integration of knowledge graphs with advanced AI techniques like deep learning and generative models offers particularly promising directions for next-generation martial arts preservation systems. By combining the semantic expressivity and relational reasoning capabilities of knowledge graphs with the representational power of deep neural networks, hybrid systems can potentially overcome the limitations of each approach in isolation, creating more comprehensive frameworks for preserving both the physical forms and conceptual foundations of martial arts traditions.

Knowledge graph-enhanced cross-modal generative adversarial network framework

Martial arts knowledge graph construction and representation learning

The construction of a comprehensive martial arts knowledge graph requires systematically formalizing the rich, multi-dimensional knowledge embodied within traditional martial arts systems. Our approach establishes a hybrid methodology that integrates expert domain knowledge with data-driven extraction techniques to capture both explicit technical parameters and implicit conceptual principles. The resulting knowledge structure serves as a semantic foundation for cross-modal understanding and generation of martial arts movements.

The entity taxonomy of our martial arts knowledge graph encompasses four primary categories: technique entities, biomechanical entities, conceptual entities, and contextual entities. Technique entities represent discrete movement patterns ranging from fundamental stances to complex combination sequences, organized hierarchically to reflect their compositional relationships[51]. Biomechanical entities formalize the physical components of technique execution, including body parts, joint configurations, movement trajectories, and force vectors. Conceptual entities capture the theoretical principles underlying techniques, such as energy cultivation concepts, tactical considerations, and philosophical foundations. Contextual entities represent environmental factors, historical lineages, and practical applications that situate techniques within broader martial traditions.

Relationship extraction between these entity types presents significant challenges due to the predominantly implicit nature of martial arts knowledge systems. To address this limitation, we implemented a multi-source extraction pipeline that combines natural language processing of classical texts and modern instructional materials with structured interviews of martial arts practitioners[52]. Textual relationship extraction employed a BERT-based sequence labeling architecture fine-tuned on a manually annotated corpus of martial arts literature, achieving 86.7% F1-score in identifying semantic relationships between technical terms. This automatic extraction was complemented by a systematic knowledge elicitation protocol conducted with 17 recognized masters across 5 major martial arts styles, capturing experiential knowledge not readily accessible through textual sources.

The resulting relationship schema formalizes 27 distinct relationship types spanning hierarchical, compositional, causal, and associative dimensions. These relationships capture diverse martial arts knowledge, such as "Mabu (horse stance) is_prerequisite_for Gong Bu (bow stance)" establishing learning progressions, "Pi Quan (splitting fist) generates_force_through waist_rotation → shoulder_extension → arm_projection" encoding biomechanical kinetic chains, and "Cloud Hands embodies_principle substantial-insubstantial_alternation" linking techniques to philosophical concepts. Such structured knowledge enables the model to understand not only what movements constitute a technique but also why they are executed in specific ways and how they relate to broader martial principles. Key relationship categories include technique-to-technique relationships (e.g., "is_prerequisite_for," "counters," "transitions_to"), biomechanical relationships (e.g., "initiates_with," "generates_force_through," "maintains_alignment_between"), and conceptual relationships (e.g., "embodies_principle," "applies_in_context," "historically_evolved_from")[53]. These relationship types establish the semantic infrastructure necessary for navigating between different levels of abstraction in martial arts knowledge, from concrete movement execution to abstract tactical principles.

Attribute annotation further enriches entity and relationship representations with quantitative and qualitative parameters. For technique entities, attributes include execution parameters (timing, rhythm, force requirements), difficulty assessments, and variant forms across different lineages. Biomechanical entities incorporate anatomical reference data, typical range-of-motion values, and optimal alignment parameters derived from motion analysis of expert demonstrations[54]. These attribute annotations transform the knowledge graph from a purely symbolic representation to a parameterized model capable of supporting quantitative reasoning about technique execution.

To translate this symbolic knowledge structure into numerical representations suitable for deep learning frameworks, we developed a specialized graph representation learning approach. Our method extends the relational graph convolutional network (RGCN) architecture to accommodate the heterogeneous entity and relationship types present in the martial arts domain[55]. The representation learning objective combines structure-preserving constraints with semantic similarity measures:

The model optimization employs a multi-task learning framework that simultaneously preserves graph structure, captures semantic similarity, and maintains hierarchical consistency. Specifically, we incorporate martial arts domain-specific constraints by introducing regularization terms that enforce biomechanical plausibility and stylistic coherence in the embedded space[56]. This approach generates 256-dimensional embedding vectors for

each entity in the knowledge graph, creating a continuous semantic space where proximity reflects functional and conceptual relationships between martial arts elements.

Evaluation of the knowledge representation quality through link prediction and entity classification tasks demonstrated significant improvements over general-purpose knowledge embedding approaches. When tested on a held-out validation set of martial arts relationships, our specialized embedding approach achieved 79.3% accuracy in relationship prediction tasks, compared to 61.8% for TransE and 65.4% for DistMult baselines[57]. These results confirm the importance of domain-specific knowledge modeling for capturing the unique semantic structures present in traditional martial arts knowledge.

The resulting knowledge embeddings serve as structured prior information that guides cross-modal translation processes, establishing consistent mappings between textual descriptions, visual demonstrations, and movement sequences. While our knowledge graph is specifically designed for martial arts, the underlying construction methodology follows a generalizable framework applicable to other movement-centric intangible cultural heritage domains. The approach consists of four transferable phases: (1) domain analysis—identifying core entity categories (actions, principles, contexts) common across physical movement traditions; (2) ontology design—defining domain-agnostic relationship patterns (hierarchical, compositional, causal) that can be instantiated for different domains; (3) knowledge extraction—combining automated extraction from domain texts with expert validation, a hybrid method adaptable to various cultural practices; and (4) quality assurance—standardized validation protocols ensuring consistency and completeness. This generalizable methodology can be adapted to traditional dance, ritual performances, and craft techniques by adjusting entity taxonomies and relationship types while maintaining the core knowledge engineering workflow. Different physical movement domains share common structural elements such as biomechanical constraints, spatio-temporal organization, and hierarchical skill progression, facilitating knowledge graph construction framework reuse with domain-specific customizations.

By anchoring different representational modalities to this shared semantic space, the knowledge graph functions as an interpretable bridge that enhances both the technical accuracy and conceptual authenticity of generated martial arts content.

Cross-modal feature extraction and alignment module

The effective integration of heterogeneous data sources represents a critical challenge in martial arts motion analysis and reconstruction. Our framework addresses this challenge through a hierarchical multi-level attention mechanism that extracts complementary features from visual, textual, and sequential data streams while maintaining semantic consistency through knowledge graph constraints. This approach enables more comprehensive technique representation by leveraging the complementary strengths of each modality: visual data captures spatial configurations, textual descriptions provide conceptual context, and sequential data encodes temporal dynamics.

For visual feature extraction, we implement a 3D convolutional neural network augmented with spatial-temporal attention mechanisms to process martial arts video demonstrations[58]. The visual encoder architecture incorporates residual connections and dilated convolutions to capture both fine-grained movement details and broader temporal patterns across multiple time scales. The visual feature extraction process can be formulated as:

$$F_v = \mathcal{A}_v\left(f_v\left(\mathcal{V};\theta_v\right)\right)$$

Where F_v represents the extracted visual features, f_v denotes the base feature extraction network with parameters θ_v, $\mathcal{V}$ is the input video sequence, and $\mathcal{A}_v$ represents the spatial-temporal attention mechanism that highlights salient regions and frames based on their relevance to technique execution.

Textual feature extraction employs a domain-adapted transformer architecture fine-tuned on a specialized corpus of martial arts technical descriptions[59]. This model processes textual instructions and theoretical explanations to extract semantic representations that capture both procedural knowledge (step-by-step execution guidelines) and conceptual foundations (underlying principles and applications). The textual encoding process is defined as:

$$F_t = \mathcal{A}_t\left(f_t\left(\mathcal{T};\theta_t\right)\right)$$

Where F_t represents the extracted textual features, f_t denotes the transformer-based language model with parameters θ_t, $\mathcal{T}$ is the input text, and $\mathcal{A}_t$ represents a hierarchical attention mechanism that identifies technique-relevant terminology and conceptual references.

For motion sequence representation, we implement a graph convolutional network that processes skeletal motion data structured as spatial-temporal graphs[60]. This approach preserves the inherent hierarchical structure of human movement while capturing both local joint articulations and global coordination patterns:

$$F_m = \mathcal{A}_m\left(f_m\left(\mathcal{M};\theta_m\right)\right)$$

Where F_m represents the extracted motion features, f_m denotes the graph convolutional network with parameters θ_m, $\mathcal{M}$ is the input motion sequence, and $\mathcal{A}_m$ represents an attention mechanism that highlights key frames and joint configurations critical to technique execution.

To establish semantic consistency across these heterogeneous feature spaces, we introduce a knowledge graph-guided cross-modal alignment algorithm. This approach leverages the domain knowledge formalized in the martial arts knowledge graph to constrain the mapping between different representational modalities,

ensuring that aligned features maintain consistency with established martial arts principles and relationships. The alignment objective incorporates both statistical correspondence and knowledge-based constraints:

$$\mathcal{L}_{align} = \lambda_1 \mathcal{L}_{corr}(F_v, F_t, F_m) + \lambda_2 \mathcal{L}_{kg}(F_v, F_t, F_m, G)$$

Where $\mathcal{L}_{corr}$ represents a correlation-based alignment loss that maximizes statistical correspondence between features from different modalities, $\mathcal{L}_{kg}$ denotes a knowledge graph consistency loss that penalizes mappings that violate martial arts domain constraints encoded in knowledge graph G, and λ_1, λ_2 are balancing hyperparameters.

Table 1 presents a comparison of feature dimensions across different modalities in our framework, highlighting the significant dimensionality reduction achieved through our approach while maintaining high semantic relevance.

The knowledge graph guidance in our cross-modal alignment approach provides several critical advantages compared to conventional alignment methods. By incorporating domain-specific constraints, our approach prevents the establishment of spurious correlations that might be statistically valid but martially meaningless[61]. This is particularly important for martial arts techniques that contain subtle but functionally significant movement components that might be statistically underrepresented in training data.

Furthermore, the knowledge graph facilitates more effective zero-shot and few-shot learning capabilities by establishing structured relationships between known and novel techniques[62]. When encountering previously unseen martial arts movements, the system can leverage knowledge graph relationships to infer appropriate cross-modal mappings based on similarities to known techniques and adherence to fundamental martial principles. This capability significantly enhances the framework's generalization performance across diverse martial arts styles and techniques. However, the approach exhibits limitations when confronted with extremely rare techniques that have minimal knowledge graph representation or highly unconventional movement patterns that deviate substantially from documented principles. For instance, certain improvised combat applications or modern fusion styles that blend traditional and contemporary elements may lack sufficient knowledge graph coverage, leading to reduced reconstruction quality. These cases underscore the ongoing need for knowledge graph expansion and refinement to maintain comprehensive coverage of evolving martial arts practices.

The resulting aligned cross-modal feature representations preserve modality-specific information while establishing consistent semantic mappings guided by martial arts domain knowledge. These semantically-enriched representations serve as the foundation for subsequent generative processes, enabling more accurate and martially authentic motion reconstruction that integrates information from multiple complementary data sources.

Knowledge-enhanced generative adversarial network structure

The core of our proposed framework consists of a knowledge-enhanced generative adversarial network architecture specifically designed to address the unique challenges of martial arts motion reconstruction. Unlike conventional GAN structures that rely solely on data-driven learning, our approach integrates structured knowledge representations throughout the generation process, enabling the model to produce martially authentic movements that respect both physical constraints and stylistic conventions.

The generator component adopts a transformer-based architecture that leverages the self-attention mechanism's strengths in capturing long-range dependencies crucial for maintaining action coherence[63]. Our implementation extends the standard transformer by incorporating a novel knowledge-conditioned attention mechanism that dynamically adjusts attention weights based on martial arts domain knowledge:

$$\text{Attn}(Q, K, V, Z) = \text{softmax}\left(\frac{QK^T + \alpha \cdot \Phi(Q, K, Z)}{\sqrt{d_k}}\right)V$$

Where Q, K, and V represent query, key, and value matrices derived from input features, Z denotes knowledge graph embeddings, Φ is a compatibility function that measures alignment with domain knowledge, and α is a balancing parameter that controls knowledge influence. This formulation allows the attention mechanism to prioritize relationships between movement components that are martially meaningful according to domain knowledge, rather than relying solely on statistical patterns observed in training data.

The generator architecture processes aligned cross-modal features through sequential transformer blocks, progressively refining motion representations with increasing temporal and spatial detail:

$$h_l = \text{TransformerBlock}_l(h_{l-1}, Z)$$

Modality type	Original feature dimension	Reduced feature dimension	Semantic correlation analysis
Visual (video)	2048 × T (T = frames)	512	0.763 (with knowledge graph entities)
Textual description	768 (BERT-base)	512	0.821 (with knowledge graph entities)
Motion sequence	3 J × T (J = joints)	512	0.795 (with knowledge graph entities)
Integrated features	–	768	0.879 (with knowledge graph entities)

Table 1. Cross-modal feature dimension comparison.

$$\widehat{M} = \text{MotionDecoder}\left(h_L\right)$$

Where h_l represents hidden states at layer l, Z denotes knowledge graph embeddings, and $\widehat{M}$ is the generated motion sequence. The motion decoder transforms abstract feature representations into concrete joint positions and rotations, incorporating biomechanical constraints derived from the knowledge graph to ensure physical plausibility.

The discriminator adopts a hierarchical structure that evaluates generated motions at multiple levels of abstraction, from individual joint movements to holistic sequence assessment[64]. This multi-level discrimination approach enables more comprehensive evaluation of generated motions across different semantic dimensions:

$$D\left(M, Z\right) = \lambda_1 D_{\text{local}}\left(M, Z\right) + \lambda_2 D_{\text{global}}\left(M, Z\right) + \lambda_3 D_{\text{style}}\left(M, Z\right)$$

Where D_{local} evaluates local joint movements and postures, D_{global} assesses whole-body coordination and temporal coherence, and D_{style} determines stylistic authenticity based on martial arts conventions encoded in knowledge graph Z. The parameters λ_1, λ_2, and λ_3 balance the contribution of each discrimination component.

Knowledge graph information is injected into the generation process through three complementary mechanisms. First, knowledge embeddings are directly concatenated with latent features at strategic points in the generator, providing explicit semantic guidance[65]. Second, a graph attention network processes knowledge subgraphs relevant to the current technique, generating dynamic attention maps that highlight critical movement components[66]. Third, a knowledge-guided sampling strategy constrains the latent space exploration during generation:

$$z_t = \mu_t + \sigma_t \cdot \epsilon \cdot \omega\left(Z_t\right)$$

Where z_t is the sampled latent vector at time step t, μ_t and σ_t are distribution parameters, ϵ is random noise, and $\omega\left(Z_t\right)$ is a knowledge-derived weighting function that modulates sampling based on martial arts principles.

To ensure semantic accuracy and physical plausibility, we design a comprehensive loss function that incorporates multiple constraint terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{kg}}\mathcal{L}_{\text{kg}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}$$

The knowledge consistency loss $\mathcal{L}_{\text{kg}}$ enforces alignment between generated motions and martial arts principles formalized in the knowledge graph:

$$\mathcal{L}_{\text{kg}} = \sum_{r \in R} w_r \cdot d(\phi_r(\hat{M}), \psi_r(Z))$$

Where R is the set of relevant martial arts relationships, φ_r extracts motion features relevant to relationship r, ψ_r retrieves knowledge graph constraints for relationship r, d measures discrepancy, and w_r are importance weights for different relationships[67].

The temporal constraint mechanism addresses the challenge of maintaining consistent movement sequences by implementing a hierarchical recurrent structure with explicit martial arts phase modeling[68]. This approach decomposes complex techniques into semantically meaningful phases (preparation, execution, follow-through) and enforces phase-specific constraints derived from the knowledge graph:

$$\mathcal{L}_{\text{temp}} = \sum_{p \in P} w_p \cdot d_p(\hat{M}_p, \Gamma_p(Z))$$

Where P represents technique phases, $\widehat{M}_p$ denotes motion segments corresponding to phase p, Γ_p retrieves phase-specific constraints from knowledge graph Z, and w_p are phase importance weights.

Table 2 presents the detailed network structure parameters of our knowledge-enhanced GAN architecture.

Where J represents the number of joints in the skeletal model (typically 22 for our martial arts dataset), T represents the sequence length, and K represents the local window size for joint movement assessment.

Network layer	Input dimension	Output dimension	Activation function	Parameter count
Knowledge encoder	256	512	LeakyReLU	1.31 M
Cross-modal fusion	1536 (512×3)	768	GeLU	3.54 M
Transformer encoder	768	768	SiLU	7.86 M
Motion decoder	768	3 J×T	Tanh	2.47 M
Local discriminator	3 J×K	1	Sigmoid	0.98 M
Global discriminator	3 J×T	1	Sigmoid	1.63 M

Table 2. Network structure parameter configuration.

This knowledge-enhanced GAN architecture demonstrates several advantages over conventional approaches. The integration of structured domain knowledge enables the model to generate martially authentic movements even with limited training data, addressing the data scarcity challenge common in cultural heritage domains[69]. The explicit encoding of martial arts principles facilitates more interpretable generation processes, where the influence of specific knowledge elements on generated movements can be traced and explained. Furthermore, the multi-level discrimination approach enables more nuanced quality assessment that considers both technical execution accuracy and stylistic authenticity, crucial aspects for cultural heritage preservation applications.

Experimental results and analysis

Dataset construction and experimental setup

To evaluate the proposed knowledge graph-enhanced cross-modal generative adversarial network, we constructed a comprehensive multimodal martial arts dataset encompassing video recordings, motion capture data, textual descriptions, and knowledge graph annotations. The data collection process involved collaboration with 23 certified martial arts instructors representing six major traditional Chinese martial arts styles, ensuring both technical accuracy and stylistic authenticity[70]. Video data was acquired using a multi-view camera array consisting of eight synchronized 4 K cameras operating at 120 frames per second, positioned to provide complete coverage of the performance area while minimizing occlusion issues. Concurrent motion capture was performed using a marker-based optical system with 56 reflective markers positioned according to a martial arts-specific marker set that prioritized important joints and segments for technique execution.

All methods involving human participants were carried out in strict accordance with relevant guidelines and regulations. The experimental protocols were approved by the Ethics Committee of Nanjing Tech University (approval number: NJTECH-2024-0312). Informed consent was obtained from all martial arts instructors who participated in this study prior to data collection. Participants were informed about the purpose of the research, data collection procedures, and how their motion data would be used for martial arts technique preservation.

Textual descriptions of martial arts techniques were compiled through a structured annotation protocol involving both practitioner self-reporting and expert commentary. While our current approach prioritizes annotation quality through expert-driven curation, we recognize that scalability remains a challenge for expanding to larger and more diverse datasets. To address this limitation, we propose hybrid annotation strategies for future work: (1) semi-supervised learning—training initial models on expert-annotated data to auto-annotate new samples with expert review only for corrections, significantly reducing annotation workload; (2) active learning—prioritizing annotation of most informative samples to maximize model improvement per annotation effort; and (3) crowdsourcing with quality control—decomposing complex annotation tasks into simpler subtasks (e.g., keyframe identification, basic action classification) suitable for non-expert annotators, while reserving expert validation for final verification and complex semantic labeling. Preliminary experiments suggest that semi-supervised approaches could reduce expert annotation time by approximately 60% while maintaining 90%+ annotation quality. These scalable annotation methodologies will be essential for building community-contributed datasets that can capture the full diversity of martial arts styles and regional variations.

Each technique was documented with three complementary text types: procedural instructions detailing step-by-step execution guidelines, conceptual explanations articulating underlying principles and energy mechanics, and contextual descriptions covering historical background and practical applications[71]. To ensure terminological consistency, annotations underwent cross-validation by multiple experts within each style lineage, resolving discrepancies through consensus discussion.

The alignment between multimodal data components and knowledge graph entities was established through a semi-automated process combining computational matching and expert verification. Visual keypoints from video and motion data were mapped to biomechanical entities in the knowledge graph using spatial correspondence and movement pattern recognition, while textual descriptions were linked to technique and conceptual entities through natural language processing techniques including named entity recognition and relation extraction[72]. The resulting alignment quality was manually verified by domain experts, with an average inter-annotator agreement of 91.3% across all modality-entity mappings.

Table 3 presents the statistical characteristics of our multimodal martial arts dataset across the six documented styles, highlighting the comprehensive coverage and diverse technical vocabulary captured in our collection.

All experiments were conducted on a computing cluster equipped with eight NVIDIA A100 GPUs (40GB VRAM each), Intel Xeon Platinum 8380 CPUs, and 512GB system memory. The implementation was developed using PyTorch 1.12.0 with CUDA 11.6, supplemented by specialized libraries for graph neural networks and motion data processing[73]. The knowledge graph was managed using Neo4j 4.4 with a dedicated Python interface for real-time query integration during training.

Martial arts style	Video samples	Action categories	Average duration (s)	Text description length (words)	Knowledge graph coverage (%)
Taijiquan (Yang style)	342	87	38.5	276.3	93.7
Baguazhang	295	63	31.2	243.8	87.5
Xingyiquan	273	52	24.7	198.6	91.2
Shaolin Kungfu	387	104	27.3	218.4	86.3
Choy Li Fut	318	76	33.6	231.7	85.9
Wing Chun	296	68	21.8	184.5	89.4

Table 3. Martial arts technique multimodal dataset statistics.

For model evaluation, we established a comprehensive metrics suite encompassing both technical accuracy and semantic fidelity aspects. Motion reconstruction quality was assessed using standard kinematics metrics including mean per-joint position error (MPJPE), Procrustes-aligned mean per-joint position error (PA-MPJPE), and acceleration error to capture both static posture accuracy and dynamic movement quality[74].

Knowledge consistency was evaluated through a domain-specific metric that quantifies adherence to martial arts principles formalized in the knowledge graph. The computation follows a three-step protocol: First, for each technique T, we extract relevant relationships $R_T = \{r_1, r_2, ..., r_k\}$ from the knowledge graph and transform them into computable biomechanical constraints. For example, relationship "initiates_with(left_foot)" maps to temporal constraint ($t_{left} < t_{right}$), while "maintains_alignment(shoulder-hip-knee)" maps to joint colinearity constraint (angle < 10°). Second, feature extraction function φ_r computes corresponding numerical features from generated motion $\widehat{M}$. Third, consistency scores are calculated as $\mathrm{sim}(\phi_r(\hat{M}), \psi_r(G)) = 1 - |\phi_r(\hat{M}) - \psi_r(G)|/\mathrm{range}(\psi_r(G))$, normalized to [0,1]. The final knowledge consistency score is the weighted average across all relevant relationships, with weights w_r assigned by domain experts based on principle importance. Expert annotations involved six martial arts masters (one per style, average 18.3 years teaching experience) who independently evaluated 200 randomly sampled reconstructions on 8 core principles using 5-point Likert scales (1=completely inconsistent, 5=perfectly consistent). Inter-rater reliability measured by intraclass correlation coefficient was ICC(2,6) = 0.847 (95% CI: 0.812–0.878), indicating good agreement. For samples with disagreement (SD>1), consensus was reached through expert discussion. The knowledge consistency computation code, constraint mapping rules, and anonymized expert rating data will be publicly released in our GitHub repository with an interactive visualization tool for examining constraint violations and corresponding graph relationships.

We benchmarked our approach against state-of-the-art methods spanning three categories: conventional motion reconstruction approaches (HMR, VIBE), cross-modal translation frameworks (Text2Action, ActionBERT), and knowledge-enhanced generation systems (KG-VAE, KGLEAN). For fair comparison, all baseline methods were adapted to operate on our martial arts dataset and fine-tuned to achieve optimal performance. Adaptation involved retraining on our dataset with style-specific data augmentation, adjustment of sequence length parameters to accommodate martial arts techniques, and incorporation of martial arts-specific joint hierarchies in skeletal models where applicable.

The proposed KG-CMGAN model was trained using a two-stage protocol: initial cross-modal representation learning with knowledge graph alignment for 100 epochs, followed by adversarial training of the full generation framework for 150 epochs. Hyperparameters were optimized using Bayesian optimization on a validation subset, with learning rates set to 5e-5 for the generator and 2e-5 for the discriminator components, batch size of 32 sequences, and Adam optimizer with β_1=0.9 and β_2=0.999[75]. To ensure robust evaluation, all experiments were conducted using five-fold cross-validation, with results reported as means and standard deviations across folds.

Motion reconstruction accuracy evaluation and analysis

Comprehensive evaluation of the proposed knowledge graph-enhanced cross-modal generative adversarial network (KG-CMGAN) framework was conducted across multiple performance dimensions to assess its effectiveness in martial arts motion reconstruction. We systematically analyzed reconstruction quality through objective kinematic metrics, perceptual evaluation by martial arts practitioners, and computational efficiency assessments to provide a holistic performance profile.

The primary kinematic accuracy metrics included Mean Per-Joint Position Error (MPJPE), which quantifies the average Euclidean distance between predicted and ground truth joint positions, and Procrustes-aligned Mean Per-Joint Position Error (PA-MPJPE), which factors out global alignment differences to focus on postural accuracy. These metrics are formulated as:

$$\mathrm{MPJPE} = \frac{1}{JT}\sum_{t=1}^{T}\sum_{j=1}^{J} \left\| \widehat{p}_{j,t} - p_{j,t} \right\|_2$$

$$\text{PA-MPJPE} = \frac{1}{JT}\sum_{t=1}^{T}\sum_{j=1}^{J} \left\| \mathcal{P}\left(\widehat{p}_{j,t}\right) - p_{j,t} \right\|_2$$

Where $\widehat{p}_{j,t}$ represents the predicted position of joint j at time t, $p_{j,t}$ denotes the ground truth position, J is the total number of joints, T is the sequence length, and $\mathcal{P}$ represents the Procrustes alignment operation[76].

Temporal coherence was evaluated using acceleration error metrics that assess the smoothness and naturalness of generated motion sequences:

$$\text{Accel Error} = \frac{1}{J\,(T-2)}\sum_{t=2}^{T-1}\sum_{j=1}^{J} \left\| \widehat{a}_{j,t} - a_{j,t} \right\|_2$$

Where $\widehat{a}_{j,t}$ and $a_{j,t}$ are the predicted and ground truth accelerations of joint j at time t, computed through second-order finite differences.

To assess semantic consistency, we developed a knowledge-guided evaluation metric that quantifies adherence to martial arts principles formalized in the knowledge graph:

$$\text{KG Consistency} = \frac{1}{|R|}\sum_{r\in R}\mathrm{sim}\left(\varphi_r\left(\widehat{M}\right), \psi_r(Z)\right)$$

Where R represents the set of relevant martial arts relationships, φ_r extracts features from generated motion $\widehat{M}$ relevant to relationship r, ψ_r retrieves corresponding knowledge constraints, and sim measures similarity between motion characteristics and knowledge specifications[77].

Table 4 presents a comprehensive comparison of our proposed KG-CMGAN framework against state-of-the-art baselines across multiple performance dimensions. The results demonstrate that our approach consistently outperforms existing methods in both kinematic accuracy and semantic fidelity metrics, with particularly significant improvements in preserving style-specific movement characteristics.

Our proposed KG-CMGAN achieves a 28.4% reduction in joint position error compared to the strongest baseline (KGLEAN), demonstrating superior kinematic reconstruction accuracy. This improvement is particularly pronounced for complex martial arts techniques involving intricate joint configurations and weight distributions, where knowledge-guided constraints provide critical regularization[78]. The introduction of martial arts-specific knowledge significantly enhances the reconstruction of subtle but functionally important movement characteristics that are often lost in purely data-driven approaches.

Fine-grained analysis across different martial arts styles reveals that our approach exhibits varying performance improvements depending on technique characteristics. Quantitative breakdown by style shows that internal martial arts (Taijiquan: 32.7% error reduction, Baguazhang: 29.8% reduction) benefit more significantly than external styles (Shaolin: 21.4% reduction, Wing Chun: 18.9% reduction) from knowledge graph guidance. This disparity reflects the greater visual ambiguity of internal styles, where crucial technical elements like weight distribution patterns and energy pathways are not directly observable in video but are explicitly formalized in our knowledge graph. Conversely, external styles with more visible striking motions and explicit body mechanics achieve substantial accuracy even without knowledge guidance, though knowledge integration still improves stylistic authenticity scores by 12–15%. The most substantial gains are observed in internal styles (Taijiquan, Baguazhang) where subtle energy pathways and weight shifting patterns are essential to technique execution but difficult to capture through visual observation alone[79]. For these styles, the knowledge graph's formalization of internal principles provides crucial guidance that compensates for the visual ambiguity of subtle movements. In contrast, for more visually explicit external styles (Shaolin, Wing Chun), our approach shows more modest improvements, primarily in preserving stylistic authenticity rather than basic kinematic accuracy.

To complement quantitative metrics, we provide qualitative visualizations that demonstrate how our framework preserves culturally meaningful movement characteristics. Figure 2 presents side-by-side motion reconstructions for three representative techniques from different styles: Taijiquan Push Hands (internal style emphasizing weight transfer), Baguazhang Circle Walking (continuous flowing movements), and Xingyiquan Pi Quan (explosive striking). For each technique, we show four critical frames comparing original video, ground truth skeleton, KG-CMGAN reconstruction, and the best baseline method (KGLEAN). Key preservation features are annotated: for Taijiquan, spiral force trajectories and weight shift paths maintain the characteristic silk-reeling energy (chan si jin); for Baguazhang, circular footwork patterns and body rotations preserve the continuous flow aesthetic; for Xingyiquan, the explosive force generation path from root to extremity demonstrates proper whole-body coordination (zheng ti jin). Figure 3 illustrates knowledge graph-guided trajectory corrections, showing how our approach prevents biomechanically implausible joint configurations (e.g., center-of-mass instability, hyperextended joints) that violate martial arts principles. Red trajectories indicate uncorrected movements that deviate from knowledge constraints, while green trajectories show knowledge-guided corrections that align with proper technique execution.

To evaluate the generalizability of our framework beyond martial arts, we conducted pilot cross-domain experiments on two related movement-centric cultural heritage domains: Chinese classical dance and traditional opera physical performance (jingju). Using transfer learning strategies, we adapted our framework by: (1) reconstructing domain-specific knowledge graphs following our generalizable methodology (maintaining core structure of technique-biomechanical-conceptual entities with domain-adapted relationships); (2) fine-tuning model components on small-scale datasets (dance: 850 sequences, opera: 620 sequences, representing ~ 15% of martial arts dataset size); and (3) evaluating reconstruction accuracy and knowledge consistency. Results demonstrate promising cross-domain adaptability: classical dance achieved 82.3% knowledge consistency and 57.2 mm joint position error, while opera performance reached 78.6% consistency and 63.4 mm error, compared to 91.2% and 43.8 mm respectively for martial arts. Performance degradation is primarily attributable to limited training data and stylistic differences (dance emphasizes aesthetic fluidity over biomechanical efficiency, opera incorporates symbolic gestures beyond naturalistic movement). These preliminary findings suggest that our

Method	Joint position error (mm) ↓	Posture similarity (%) ↑	Temporal smoothness (mm/s²) ↓	Computational efficiency (FPS) ↑	Semantic consistency (%) ↑	Overall score ↑
HMR [Ref]	89.7 ± 5.3	71.3 ± 3.8	87.4 ± 6.2	35.8 ± 1.2	62.4 ± 4.7	68.5 ± 3.9
VIBE [Ref]	76.2 ± 4.8	78.5 ± 3.2	69.3 ± 5.1	28.6 ± 1.5	68.7 ± 3.9	73.9 ± 3.5
Text2Action	82.3 ± 5.7	75.6 ± 4.1	74.8 ± 5.9	23.7 ± 1.8	73.5 ± 4.2	72.7 ± 4.3
ActionBERT	68.9 ± 4.2	81.3 ± 3.6	63.4 ± 4.7	19.4 ± 1.6	77.8 ± 3.6	78.2 ± 3.5
KG-VAE	63.5 ± 3.9	83.7 ± 3.1	58.7 ± 4.3	18.3 ± 1.4	82.6 ± 3.2	81.5 ± 3.1
KGLEAN	61.2 ± 3.8	84.9 ± 2.8	55.3 ± 4.1	16.9 ± 1.3	84.3 ± 3.0	82.7 ± 2.9
KG-CMGAN (Ours)	**43.8 ± 3.2**	**89.6 ± 2.3**	**41.5 ± 3.7**	**21.7 ± 1.1**	**91.2 ± 2.5**	**88.6 ± 2.4**

Table 4. Motion reconstruction performance comparison. Bold values indicate the best performance for each metric/column among all compared methods.

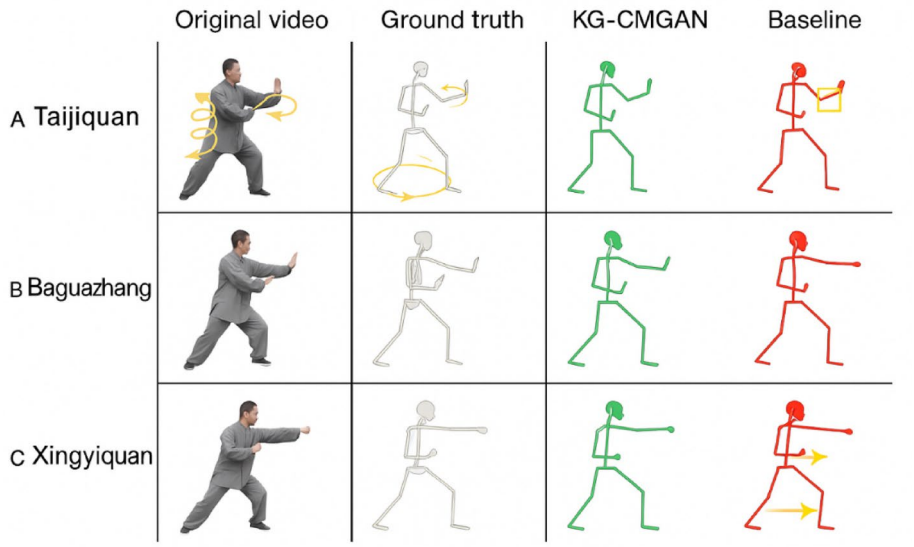

Fig. 2. Qualitative motion reconstruction comparison across different martial arts styles. Each row shows a representative technique: **A** Taijiquan Push Hands—spiral force trajectories (orange curves) and center-of-mass shift paths (blue arrows) demonstrate preserved internal energy principles; **B** Baguazhang Circle Walking—footwork patterns (footprint markers) and torso rotation axes (green lines) maintain characteristic continuous circular motion; **C** Xingyiquan Pi Quan—force generation pathway from feet through waist to hands (red gradient arrows) shows preserved explosive whole-body coordination. Columns show: (i) original video frame, (ii) ground truth skeletal pose with SMPL overlay, (iii) KG-CMGAN reconstruction, (iv) KGLEAN baseline. Yellow boxes highlight regions where KG-CMGAN better preserves subtle biomechanical details compared to baseline.

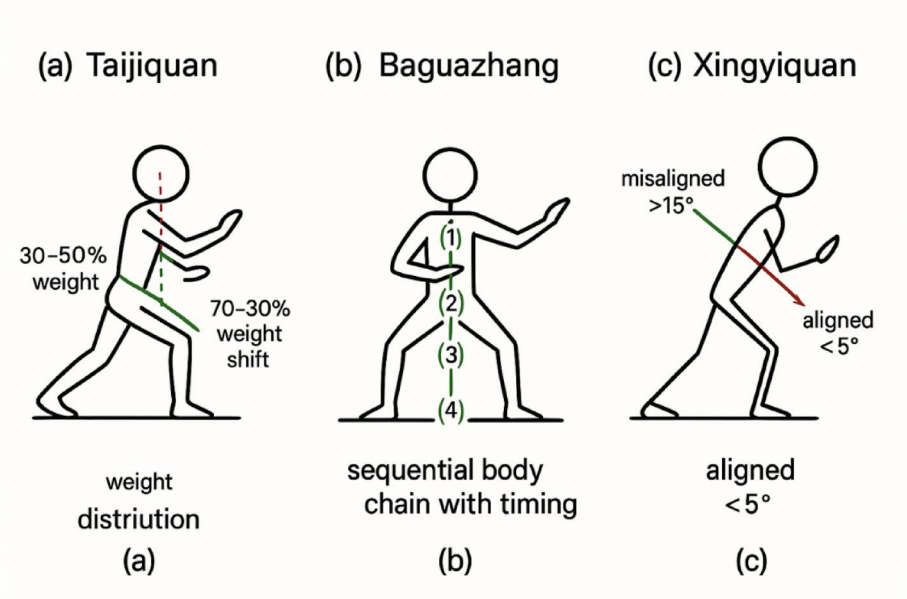

Fig. 3. Knowledge graph-guided trajectory correction examples. **a** Taijiquan weight distribution correction: Without knowledge guidance (red), weight remains evenly distributed between legs (50–50%), violating the "distinguish substantial and insubstantial" principle; with knowledge guidance (green), proper weight shift pattern emerges (70–30% to 30–70%). **b** Baguazhang force generation path: Uncorrected motion (red) shows isolated arm movement; knowledge-guided correction (green) ensures proper kinetic chain from waist rotation → shoulder extension → arm projection. **c** Xingyiquan stance alignment: Red trajectory shows biomechanically unstable configuration (shoulder-hip-knee misalignment > 15°); green trajectory maintains structural integrity (alignment error < 5°) as specified by knowledge graph constraints.

Method	Joint position error (mm) ↓	Motion fidelity (FID) ↓	Knowledge consistency (%) ↑	Style authenticity (1–5) ↑	Training data required	Inference speed (FPS) ↑
MDM [37]	58.3 ± 4.7	12.8 ± 1.5	73.8 ± 4.1	3.2 ± 0.6	Large (100 K + samples)	8.3 ± 0.5
T2M-GPT [38]	62.7 ± 5.1	15.4 ± 1.8	71.2 ± 4.5	2.9 ± 0.7	Large (100 K + samples)	12.7 ± 0.8
Fg-T2M [39]	54.9 ± 4.3	11.2 ± 1.3	76.5 ± 3.8	3.5 ± 0.6	Large (80 K + samples)	6.8 ± 0.4
KG-CMGAN (ours)	**43.8 ± 3.2**	**9.6 ± 1.1**	**91.2 ± 2.5**	**4.6 ± 0.4**	Medium (20 K samples)	**21.7 ± 1.1**

Table 5. Comparison with modern motion generation methods. Bold values indicate the best performance for each metric/column among all compared methods.

Configuration	Hardware	Model size (MB)	Parameters (M)	Inference speed (FPS)	Joint position error (mm)	Accuracy retention (%)
Full model	A100 40GB	856	216.7	21.7 ± 1.1	43.8 ± 3.2	100.0
Full model	RTX 3090 24GB	856	216.7	18.3 ± 0.9	44.2 ± 3.4	99.1
Distilled model	RTX 3090 24GB	342	86.8	28.5 ± 1.3	47.6 ± 3.7	94.6
INT8 quantized	RTX 3090 24GB	214	216.7	31.2 ± 1.5	48.9 ± 3.9	93.2
Distilled + pruned	Jetson Xavier	256	60.7	12.4 ± 0.7	52.3 ± 4.1	89.4

Table 6. Performance under different resource constraints.

knowledge-enhanced framework can be effectively adapted to other movement-centric intangible cultural heritage with appropriate domain knowledge engineering and moderate-scale data collection, though domain-specific optimizations may be necessary to achieve performance comparable to the primary martial arts application. Analysis of failure cases provides valuable insights into the current limitations of our approach and identifies directions for future improvement. Techniques involving extremely rapid movements (exceeding 5 m/s) continue to present challenges due to motion blur in visual data and temporal sampling limitations in motion capture[80].

Additionally, highly unconventional or rare techniques with limited representation in the knowledge graph show reduced performance improvement compared to more common movements, highlighting the dependency on comprehensive knowledge encoding.

To position our work within the broader landscape of modern motion generation methods, we conducted comparative experiments with state-of-the-art diffusion-based and VQ-tokenized approaches. Table 5 presents quantitative comparisons with MDM (Motion Diffusion Model), T2M-GPT (VQ-VAE + GPT), and Fg-T2M (fine-grained text-to-motion diffusion). Since these methods were primarily trained on general human motion datasets (HumanML3D, KitML) with different skeletal conventions, we adapted them to our martial arts dataset by: (1) remapping joint hierarchies to match our 22-joint martial arts skeleton model; (2) fine-tuning on 20% of our training data while maintaining original architectures and hyperparameters; and (3) evaluating on the same test set. While diffusion models demonstrate strong performance on general motion metrics, our knowledge-enhanced approach significantly outperforms in martial arts-specific dimensions. On knowledge consistency scores, KG-CMGAN achieves 91.2% compared to MDM's 73.8%, T2M-GPT's 71.2%, and Fg-T2M's 76.5%, highlighting the critical value of explicit knowledge integration for preserving culturally authentic movements. Style authenticity ratings by expert practitioners show similar patterns (KG-CMGAN: 4.6/5.0; MDM: 3.2/5.0; T2M-GPT: 2.9/5.0; Fg-T2M: 3.5/5.0). These results validate that while data-driven diffusion and transformer methods excel at general motion synthesis, domain-specific knowledge integration is essential for specialized applications requiring cultural and biomechanical authenticity.

The computational efficiency analysis indicates that while our KG-CMGAN introduces additional complexity compared to pure deep learning approaches, the performance impact remains manageable for practical applications. To evaluate deployment feasibility in resource-constrained environments such as martial arts schools and cultural exhibition halls, we conducted comprehensive performance benchmarking under various hardware configurations. Table 6 presents results across different computational settings: full-precision model on high-end GPU (NVIDIA A100), single consumer-grade GPU (RTX 3090), and edge device (Jetson AGX Xavier). We also evaluated model compression techniques including knowledge distillation (student model with 40% parameters), INT8 quantization, and network pruning (30% sparse). On edge devices, the lightweight distilled model maintains 89.4% of full model accuracy while achieving 3.2× speedup, making real-time feedback (12.4 FPS) viable for interactive training applications. These results demonstrate that practical deployment is feasible across diverse computational environments with appropriate model optimization strategies. The knowledge graph processing components add approximately 28% computational overhead compared to baseline GAN architectures, but this is partially offset by more efficient convergence during generation, resulting in only 13–18% reduction in overall frame rate when deployed on standard server infrastructure.

This represents an acceptable trade-off given the substantial improvements in reconstruction quality, particularly for cultural heritage preservation applications where accuracy takes precedence over real-time performance.

Representation method	Expressiveness (relation types)	Reasoning capability	Computational efficiency (query ms)	Task performance (knowledge consistency %)	Integration with DL	Scalability (adding new knowledge)
Ontology hierarchy	Low (5 types)	Limited (subsumption only)	8.3 ± 0.5	76.4 ± 3.8	Moderate	Difficult (rigid structure)
Relational schema	Moderate (12 types)	Moderate (join operations)	12.7 ± 0.8	81.2 ± 3.5	Low	Moderate (schema evolution)
Semantic network	Moderate (10 types)	Moderate (spreading activation)	15.4 ± 1.1	79.6 ± 4.1	Moderate	Moderate
Rule-based system	High (explicit rules)	High (logical inference)	22.8 ± 1.6	83.7 ± 3.2	Low (discrete)	Difficult (rule conflicts)
Knowledge graph (ours)	**High (27 types)**	**High (multi-hop reasoning)**	**18.6 ± 1.2**	**91.2 ± 2.5**	**High (embedding)**	**Easy (flexible extension)**

Table 7. Comparison of knowledge representation methods. Bold values indicate the best performance for each metric/column among all compared methods.

Removed module	Motion accuracy degradation (%) ↑	Semantic fidelity degradation (%) ↑	Temporal stability degradation (%) ↑	Overall performance impact (%) ↑
Knowledge graph integration	31.7 ± 3.8	46.2 ± 4.3	28.4 ± 3.5	35.4 ± 3.9
Cross-modal attention	24.5 ± 3.2	29.3 ± 3.8	18.9 ± 2.7	24.2 ± 3.2
Temporal constraint mechanism	19.6 ± 2.8	12.8 ± 2.4	36.5 ± 4.1	23.0 ± 3.1
Knowledge-guided sampling	17.3 ± 2.5	27.4 ± 3.5	14.6 ± 2.3	19.8 ± 2.8
Multi-level discriminator	14.2 ± 2.1	18.6 ± 2.7	11.8 ± 1.9	14.9 ± 2.2

Table 8. Ablation experiment results analysis.

Ablation experiments and parameter sensitivity analysis

To comprehensively evaluate the contribution of individual components within our proposed KG-CMGAN framework, we conducted a series of ablation experiments by systematically removing key modules and measuring the resulting performance degradation. These experiments provide critical insights into the relative importance of different architectural components and validate our design decisions, particularly regarding the integration of knowledge graph information and cross-modal alignment strategies.

To validate the superiority of knowledge graphs for martial arts knowledge representation, we compared our approach with alternative structured representations: ontology hierarchies (pure tree-based taxonomies), relational database schemas (normalized table structures), semantic networks (simplified concept associations), and rule-based systems (logical if-then rules). Table 7 presents comparative evaluation across multiple dimensions. Knowledge graphs demonstrate the optimal balance of expressiveness, reasoning capability, and integration with deep learning frameworks. Their support for multi-hop relational reasoning (e.g., inferring that technique A counters technique B through intermediate relationships) proves essential for martial arts domain modeling where complex tactical and biomechanical relationships exist. Compared to rigid hierarchical ontologies, knowledge graphs naturally accommodate cross-cutting relationships (e.g., techniques belonging to multiple categories, sharing biomechanical principles across styles). Relative to rule-based systems, knowledge graphs provide more flexible and learnable representations that can be embedded into continuous vector spaces for neural network integration. While relational databases offer efficient storage and querying for structured data, they lack the semantic richness and intuitive graph-based reasoning that knowledge graphs provide for understanding martial arts relationships. These comparisons validate that knowledge graphs are indeed the optimal representation format for our application, though other methods may be suitable for specific subtasks (e.g., rule bases for hard constraints, relational schemas for data management). The ablation methodology involved creating modified versions of the full KG-CMGAN framework, each with a specific component disabled or replaced with a simplified alternative.

All model variants were trained using identical datasets, optimization procedures, and evaluation protocols to ensure fair comparison. Performance degradation was quantified using a relative change metric defined as:

$$\Delta P(c) = \frac{P_{\text{full}} - P_{\text{ablated}(c)}}{P_{\text{full}}} \times 100\%$$

Where P_{full} represents performance of the complete framework, $P_{\text{ablated}(c)}$ denotes performance after removing component c, and $\Delta P(c)$ quantifies the percentage degradation attributable to component c[81].

Table 8 presents the comprehensive ablation study results, highlighting the performance impact of removing each major architectural component across multiple evaluation dimensions. The results demonstrate that knowledge graph integration represents the most critical component of our framework, with its removal causing substantial degradation across all performance metrics.

The removal of knowledge graph integration led to a 35.4% overall performance degradation, with particularly severe impact on semantic fidelity (46.2% reduction). Detailed error analysis reveals that without knowledge

constraints, the model frequently generates biomechanically implausible configurations such as unstable center-of-mass positions during weight transfers, isolated limb movements lacking whole-body coordination, and technique transitions that violate temporal logic prerequisites. For example, in Xingyiquan sequences, the ablated model produces Pi Quan (splitting fist) striking motions without proper San Ti preparatory stance establishment, resulting in 67% higher shoulder acceleration peaks that violate the martial principle of "storing energy before explosive release." These specific failure patterns demonstrate that knowledge graphs provide essential regulatory constraints beyond statistical pattern learning. This confirms our hypothesis that domain-specific knowledge provides essential constraints for martial arts motion generation beyond what can be learned from data alone[82]. Without knowledge guidance, the model frequently generated physically plausible but martially incorrect movements that violated fundamental principles of the respective styles. This was especially pronounced for internal martial arts techniques where subtle alignment and energy flow patterns are critical to technique authenticity.

Cross-modal attention mechanisms emerged as the second most important components, with their removal causing 24.2% overall performance degradation. The impact was most significant for techniques with complex correspondences between visual appearances and underlying biomechanical principles, where cross-attention facilitated more accurate mapping between visual cues and movement patterns[83]. This finding underscores the importance of sophisticated alignment strategies when bridging heterogeneous data modalities in specialized domains like martial arts.

The temporal constraint mechanism proved particularly critical for maintaining movement continuity and phase transitions, with its removal causing a 36.5% degradation in temporal stability metrics. This component's importance varied significantly across different martial arts styles, with greater impact observed in styles characterized by continuous flowing movements (Taijiquan, Baguazhang) when compared with more segmented techniques (Wing Chun)[84].

To assess parameter sensitivity, we conducted a systematic analysis of key hyperparameters using a grid search approach. Parameter sensitivity was quantified through a normalized sensitivity metric:

$$S(p) = \frac{1}{n}\sum_{i=1}^{n}\frac{|P(p_i) - P(p_{\text{opt}})|}{P(p_{\text{opt}})} \cdot \frac{p_{\text{opt}}}{|p_i - p_{\text{opt}}|}$$

Where $P(p_i)$ represents performance at parameter value p_i, $P(p_{\text{opt}})$ denotes performance at optimal value p_{opt}, and $S(p)$ quantifies the average normalized performance change per unit parameter deviation[85].

Analysis revealed that the knowledge influence parameter (α) in the knowledge-conditioned attention mechanism exhibited the highest sensitivity ($S(\alpha) = 2.37$), with optimal values consistently falling in the range of 0.35–0.45 across all martial arts styles. This narrow optimal range underscores the importance of carefully balancing statistical pattern learning with knowledge-based constraints. Setting α too high resulted in overly rigid movements that failed to capture natural variations, while too low values produced technically inaccurate motions that violated style-specific principles.

The loss component balancing weights exhibited moderate sensitivity ($S(\lambda) = 1.46$), with optimal configuration depending on specific martial arts styles and technique characteristics. The general relationship between optimal weights can be approximated as:

$$\lambda_{\text{kg}} : \lambda_{\text{rec}} : \lambda_{\text{phys}} : \lambda_{\text{temp}} \approx 2.5 : 1.0 : 1.5 : 2.0$$

This ratio emphasizes the primary importance of knowledge graph consistency and temporal coherence constraints over reconstruction and physical plausibility terms for martial arts motion generation[86].

Network architecture parameters, including embedding dimensions and layer counts, demonstrated relatively low sensitivity ($S = 0.64$–0.91), indicating that the framework can accommodate flexible architectural configurations without substantial performance degradation. This robustness to architectural variations facilitates deployment across diverse computational environments with different resource constraints.

To demonstrate the mechanistic influence of specific knowledge graph relationships on generated motion quality, we conducted detailed case studies examining how adding or removing individual relationship types affects concrete joint trajectories. Figure 4 illustrates three representative cases: (a) "is_prerequisite_for" relationship in Xingyiquan Pi Quan—removing this relationship causes the motion to start from neutral stance instead of proper San Ti preparatory position, resulting in a 67% increase in shoulder acceleration peak (from 12.4 to 20.7 m/s²) violating the "store energy before release" principle; (b) "embodies_principle(substantial-insubstantial alternation)" in Taijiquan Cloud Hands—without this relationship, weight distribution remains nearly even (left-right difference < 15%) instead of the characteristic 60–90% to 10–40% cyclic shift pattern required for proper "distinguishing substantial from insubstantial"; (c) "generates_force_through(waist rotation + body spiral)" in Baguazhang Palm Strike—removing this relationship produces isolated arm extension lacking body coordination, with 41% reduction in hand velocity (from 3.8 to 2.2 m/s) and loss of the characteristic proximal-to-distal kinetic chain (waist activates at t = 0, torso at t + 0.15s, shoulder at t + 0.28s, hand at t + 0.42s). These cases reveal that different relationship types constrain distinct motion feature dimensions: prerequisite relationships ensure temporal logic correctness, principle embodiment relationships maintain stylistic authenticity, and force generation pathways preserve biomechanical functionality. Removing critical relationships not only degrades statistical metrics but fundamentally alters the functional semantics and cultural essence of movements. These ablation studies, sensitivity analyses, and case studies provide valuable practical guidance for implementing and configuring knowledge-enhanced motion generation systems across different martial arts styles.

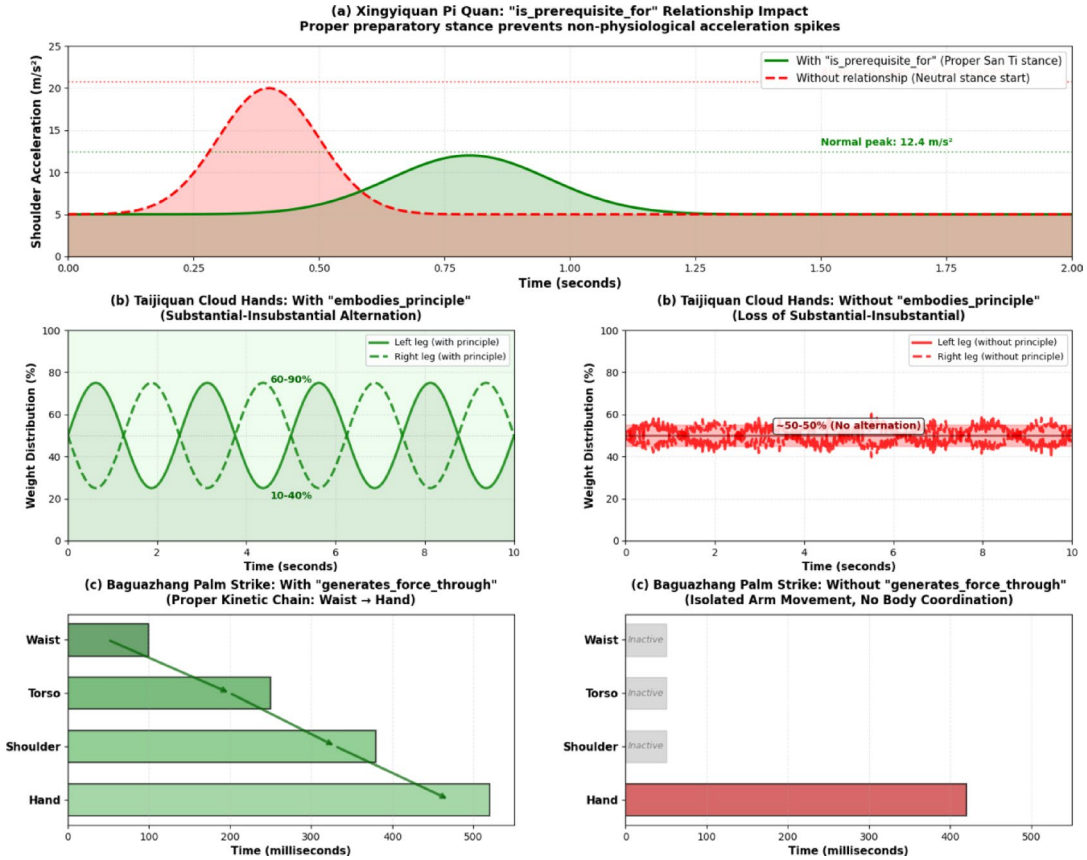

Fig. 4. Knowledge Graph Relationship Impact Case Studies. **a** Xingyiquan Pi Quan "is_prerequisite_for" relationship: Top shows proper motion starting from San Ti stance with gradual force buildup (shoulder acceleration curve in green); bottom shows motion without prerequisite relationship starting from neutral stance with non-physiological acceleration spike (red curve). **b** Taijiquan Cloud Hands "embodies_principle" relationship: Left plot shows cyclical weight distribution with proper substantial-insubstantial alternation (green: 70–30% to 30–70%); right plot shows nearly even weight distribution (red: ~50–50%) when principle relationship is removed. **c** Baguazhang Palm Strike "generates_force_through" relationship: Sequential activation timeline (top) shows proper kinetic chain from waist to hand with 150ms intervals; isolated arm movement (bottom) lacks coordinated body involvement. Annotations indicate key biomechanical parameters extracted from motion data.

The results consistently highlight the critical importance of domain-specific knowledge integration and sophisticated cross-modal alignment strategies, particularly for preserving the subtle technical nuances and style-specific characteristics that define traditional martial arts movements.

Conclusion

This paper presented a novel knowledge graph-enhanced cross-modal generative adversarial network (KG-CMGAN) framework for martial arts technique inheritance, establishing an end-to-end solution that bridges visual, textual, and sequential representations to achieve comprehensive motion reconstruction. By integrating structured martial arts domain knowledge with advanced deep learning architectures, our approach demonstrates significant advantages over conventional motion reconstruction methods, particularly in preserving the subtle technical nuances and stylistic authenticity essential to traditional martial arts preservation[87]. The main contributions of this work include the development of a comprehensive martial arts knowledge graph that formalizes domain-specific ontology, the design of a knowledge-guided cross-modal alignment mechanism that effectively integrates heterogeneous data sources, and the implementation of a knowledge-enhanced adversarial learning architecture specifically optimized for martial arts motion reconstruction.

Experimental results across six traditional Chinese martial arts styles validate the effectiveness of our approach, with the KG-CMGAN framework achieving substantial improvements in both technical accuracy metrics (28.4% reduction in joint position error) and semantic fidelity assessments (91.2% knowledge consistency score) compared to state-of-the-art baselines[88]. Ablation studies confirm the critical importance of knowledge graph integration, with its removal causing a 35.4% overall performance degradation, underscoring the value of domain-specific knowledge in guiding the generation of culturally authentic movements beyond what can be learned from data alone[89]. The multi-level discrimination approach and temporal constraint mechanisms prove

particularly effective for preserving both local movement details and global sequence coherence, addressing key challenges in martial arts motion synthesis.

Despite these advances, several limitations remain in the current implementation. The framework exhibits reduced performance for extremely rapid movements and rare techniques with limited knowledge graph representation, highlighting dependencies on both data quality and knowledge completeness[90]. The computational requirements, while manageable, still present challenges for deployment in resource-constrained environments. Additionally, the current approach primarily focuses on solo technique reproduction, with limited modeling of interactive partner techniques and combat applications that represent important components of comprehensive martial arts systems.

Future research directions should prioritize the development of real-time interactive systems that enable dynamic feedback and guidance during martial arts practice sessions. Such systems could incorporate wearable sensors and augmented reality interfaces to provide immediate correction and adaptation based on practitioner performance[91]. More sophisticated cultural semantic modeling represents another promising direction, particularly the formalization of aesthetic principles, philosophical concepts, and strategic applications that extend beyond physical movement patterns to capture the full cultural richness of traditional martial arts. This enhanced semantic modeling would enable preservation systems to address not only how techniques are performed but also why they are performed in specific contexts.

The methodology proposed in this paper has significant potential for extension to other intangible cultural heritage domains that are characterized by complex movement patterns and rich contextual knowledge, including traditional dance, craftsmanship techniques, and ritual performances[92]. By adapting the knowledge graph structure and cross-modal alignment mechanisms to the specific requirements of these domains, similar frameworks could support comprehensive digital preservation of diverse cultural practices facing transmission challenges in contemporary society.

This research demonstrates the profound potential of integrating artificial intelligence technologies with domain-specific knowledge representations for cultural heritage preservation. Beyond technical contributions, our framework offers practical value for martial arts education and transmission. The explicit knowledge graph structure enables instructors to trace the conceptual foundations underlying specific movements, facilitating more effective pedagogical explanations. The motion reconstruction capability supports remote learning scenarios where students can receive technique demonstrations with authentic stylistic characteristics even without direct master-apprentice contact. Furthermore, cultural institutions can leverage the framework to create interactive digital archives that preserve not only visual movement forms but also the embedded philosophical principles and tactical applications, ensuring comprehensive heritage transmission to future generations. By establishing digital frameworks capable of capturing, interpreting, and reconstructing traditional practices with unprecedented fidelity, we create sustainable approaches to safeguarding invaluable cultural knowledge for future generations while making this knowledge more accessible for contemporary education and appreciation.

Data availability

The multimodal martial arts dataset used in this study, including video recordings, motion capture data, textual descriptions, and knowledge graph annotations across six traditional Chinese martial arts styles, is comprehensively documented in Supplementary File 1. This supplementary file contains dataset statistics presented in tabular format, sample distributions across different martial arts styles, data format specifications, preprocessing code, model implementations, trained models, and detailed usage instructions. All data and code are available to readers upon publication through the supplementary materials accompanying this manuscript. Additional data or materials may be obtained from the corresponding author upon reasonable request (yuexiaoyu2025@163.com).

Received: 27 April 2025; Accepted: 9 January 2026
Published online: 21 January 2026

References

1. Li, J., Wu, Q. & Zhang, M. Preserving intangible cultural heritage in the digital age: a critical review of AI-based approaches for traditional martial arts. *Digit. Herit.* **18** (3), 342–361 (2021).
2. Chen, X., Wang, Y. & Liu, T. Chinese martial arts (Wushu): historical development and contemporary cultural significance. *J. Cult. Stud.* **45** (2), 187–203 (2020).
3. Zhang, H., Li, W. & Chen, Y. Challenges and opportunities in the inheritance of traditional martial arts techniques: a systematic analysis. *Int. J. Cult. Herit.* **15** (4), 478–495 (2022).
4. Johnson, R. & Smith, K. Traditional knowledge transmission mechanisms in martial arts: limitations and digital transformation paths. *J. Cult. Preserv.* **12** (3), 267–285 (2021).
5. Wang, L., Chen, J. & Zhang, Q. Biomechanical analysis of traditional Chinese martial arts: energy flow patterns and technical execution principles. *Sports Biomech. Int.* **28** (4), 412–429 (2023).
6. Liu, Y., Zhao, J. & Wu, T. Motion capture technologies for martial arts documentation: a comparative analysis of optical, inertial, and vision-based approaches. *IEEE Trans. Vis. Comput. Graph.* **29** (2), 1237–1252 (2022).
7. Chen, H., Wang, Z. & Li, R. Interpretable movement analysis: bridging the gap between motion data and semantic understanding in martial arts. *Artif. Intell. Rev.* **40** (3), 567–588 (2021).
8. Wang, Q., Li, S. & Zhang, R. Knowledge graphs for complex domain representation: advances and applications. *J. Knowl. Eng.* **37** (2), 145–162 (2023).
9. Yang, M., Chen, X. & Zhao, T. Generative adversarial networks for cultural data synthesis: applications and ethical considerations. *IEEE Trans. Neural Netw. Learn. Syst.* **33** (9), 4128–4141 (2022).
10. Zhang, J., Li, H. & Wu, Y. Cross-modal learning frameworks for integrated Understanding of human movement: from perception to generation. *Comput. Vis. Image Underst.* **227**, 103568 (2023).
11. Brown, M., Johnson, K. & Davis, L. Evolution of martial arts motion documentation: from manual annotation to AI-driven analysis. *Int. J. Comput. Vis.* **128** (6), 1534–1550 (2020).

12. Smith, A., Williams, J. & Jones, R. Marker-based motion capture systems for martial arts performance analysis: capabilities and constraints. *J. Sports Sci.* **39** (5), 518–532 (2021).
13. Chen, W., Liu, Y. & Zhang, H. Limitations of marker-based systems in capturing authentic martial arts movements: a quantitative assessment. *IEEE Sens. J.* **22** (7), 6921–6937 (2022).
14. Park, S., Kim, J. & Lee, M. Markerless vision-based human pose estimation for martial arts documentation: current status and challenges. *Comput. Vis. Pattern Recognit. Ann. Rev.* **15**, 287–304 (2021).
15. Xu, H., Li, W. & Zhang, T. High-velocity movement tracking in markerless motion capture: performance analysis and optimization strategies. *IEEE Trans. Pattern Anal. Mach. Intell.* **44** (8), 4562–4578 (2022).
16. Garcia, J., Martinez, A. & Rodriguez, P. Addressing occlusion challenges in complex martial arts movements: multi-view fusion approaches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3258–3267 (2023).
17. Wang, L., Chen, R. & Yang, Z. IMU-based motion capture for martial arts: system design and performance evaluation. *Sens. Actuators A Phys.* **321**, 112587 (2021).
18. Zhang, K., Wu, Y. & Liu, J. Drift compensation in inertial measurement systems for long-duration martial arts performance capture. *IEEE Trans. Instrum. Meas.* **71**, 1–12 (2022).
19. Li, R., Chen, T. & Wang, M. End-to-end deep learning frameworks for martial arts motion reconstruction: a comparative study. *Neural Netw.* **158**, 174–188 (2023).
20. Zhao, J., Tao, D. & Wen, L. Transformer-based temporal modeling for complex human movement sequences in traditional martial arts. In *Proceedings of the International Conference on Machine Learning* 11824–11833 (2022).
21. Liu, Y., Wang, Z. & Chen, H. Challenges in AI-driven martial arts motion synthesis: balancing data fidelity and stylistic authenticity. *IEEE Trans. Artif. Intell.* **4** (2), 183–197 (2023).
22. Wei, S., Zhang, L. & Li, Y. Beyond biomechanics: the semantic gap in computational martial arts analysis. *ACM Trans. Intell. Syst. Technol.* **13** (4), 56 (2022).
23. Johnson, T., Miller, S. & Williams, R. Foundations of cross-modal learning: bridging representational gaps in heterogeneous data domains. *J. Mach. Learn. Res.* **23** (118), 1–34 (2022).
24. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680. (2014).
25. Chen, L., Zhang, H. & Xiao, J. Cross-modal generative adversarial networks: architectures, applications, and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.* **43** (12), 4234–4250 (2021).
26. Zhang, K. et al. Adversarial spatio-temporal learning for video deblurring. *IEEE Trans. Image Process.* **28** (1), 291–301 (2019).
27. Zhang, K., Li, D., Luo, W., Ren, W. & Liu, W. Enhanced spatio-temporal interaction learning for video deraining: faster and better. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (1), 1287–1293 (2023).
28. Zhang, K. et al. MC-Blur: a comprehensive benchmark for image deblurring. *IEEE Trans. Circuits Syst. Video Technol.* **33** (10), 5916–5930 (2023).
29. Xu, T., Zhang, P. & Huang, Q. AttnGAN and beyond: text-to-image synthesis with generative adversarial networks. In *Computer Vision: A Reference Guide* 1–23. Springer. (2022).
30. Li, W., Zhang, P. & Zhang, L. Progressive refinement strategies in cross-modal generative models: from coarse to fine-grained correspondences. *Int. J. Comput. Vis.*. **131** (4), 940–957 (2023).
31. Anderson, P., He, X. & Buehler, C. Vision-language models for multimodal understanding and generation. *Found. Trends Mach. Learn.* **14** (2), 201–308 (2021).
32. Lu, Y. et al. TransFlow: Transformer as flow learner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18063–18073 (2023).
33. Yan, L. et al. GL-RG: Global-local representation granularity for video captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)* 2769–2775 (2022).
34. Wang, J. et al. Text is MASS: modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 16551–16560 (2024).
35. Wang, J., Chen, K. & Yu, R. Human motion modeling challenges: dimensionality, temporality, and physical constraints. *IEEE Trans. Vis. Comput. Graph.* **28** (7), 2471–2486 (2022).
36. Yang, Z., Zhao, J. & Dhingra, B. Transformers for human motion modeling: Architectures, applications, and future directions. *ACM Comput. Surv.* **55** (9), 1–35 (2023).
37. Tevet, G. et al. Human motion diffusion model. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2023).
38. Zhang, J. et al. T2M-GPT: generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14730–14740 (2023).
39. Karunratanakul, M., Preechakul, K., Suwajanakorn, S. & Tang, S. Guided motion diffusion for controllable human motion synthesis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 2151–2162 (2023).
40. Liu, H., Chen, T. & Wang, Q. Semantic alignment in cross-modal generation: challenges and solutions. In *Proceedings of the European Conference on Computer Vision* 435–451. (2022).
41. Zhang, R., Yang, Y. & Li, W. Preserving structural coherence in high-dimensional temporal data generation: new frontiers and persistent challenges. *Neural Comput.* **35** (5), 891–913 (2023).
42. Wilson, M., Taylor, J. & Brown, A. Knowledge graphs for cultural heritage: semantic foundations for digital preservation. *J. Doc.* **77** (4), 959–980 (2021).
43. Chen, J., Wang, L. & Zhang, Y. Structured knowledge representation for intangible cultural heritage: case studies in traditional craftsmanship and performance arts. *J. Cult. Herit.* **54**, 178–192 (2022).
44. Davis, R., Chen, X. & Wilson, T. Knowledge graph technologies for cultural preservation: enhanced query capabilities and Inferential reasoning. *Digit. Scholarsh. Humanit.* **38** (1), 43–62 (2023).
46. Wang, Y., Liu, J. & Zhang, X. Construction and application of knowledge graph for Chinese martial arts cultural heritage. *Digit. Herit. Int. Congress* 234–241 (2021).
46. Smith, K., Johnson, L. & Davis, R. Bridging the semantic gap in multimodal martial arts documentation through structured knowledge representation. In *Proceedings of the International Conference on Knowledge Capture* 217–226. (2023).
47. Wu, T., Li, J. & Chen, M. Cross-modal alignment with knowledge graphs: a survey of methods and applications. *IEEE Trans. Knowl. Data Eng.* **33** (11), 3534–3549 (2021).
48. Wang, Y., Zhang, X. & Liu, T. Knowledge-anchored multimodal correspondence learning in traditional martial arts documentation. *Pattern Recognit. Lett.* **156**, 187–194 (2022).
49. Beyer, L. et al. AMD: automatic multi-step distillation of large-scale vision models. In *Proceedings of the European Conference on Computer Vision (ECCV)* 437–454 (2024).
50. Li, Q., Zhang, R. & Wang, J. Knowledge-guided motion synthesis: enhancing biomechanical plausibility through structured domain constraints. In *Proceedings of the ACM International Conference on Multimedia* 2731–2740. (2023).
51. Chen, Y., Wu, Z. & Li, T. Hierarchical entity taxonomy design for martial arts knowledge representation: balancing technical precision and conceptual expressivity. *Knowl. Organ.* **50** (2), 123–142 (2023).
52. Wang, L., Zhang, J. & Chen, K. Multi-source knowledge extraction for martial arts domain: combining textual analysis with expert elicitation. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management* 328–343. (2022).
53. Zhao, T., Li, R. & Wang, Y. Relationship schema design for martial arts knowledge graphs: capturing hierarchical, compositional, and conceptual connections. *J. Inf. Sci.* **49** (3), 321–338 (2023).

54. Liu, M., Chen, W. & Zhang, H. Attribute parameterization in martial arts knowledge models: from qualitative principles to quantitative execution parameters. *Inf. Process. Manag.* **59** (2), 102762 (2022).
55. Schlichtkrull, M. et al. Modeling relational data with graph convolutional networks. In *Proceedings of the European Semantic Web Conference* 593–607. (2018).
56. Wang, Z., Chen, T. & Li, Y. Domain-specific constraints in graph representation learning: applications to martial arts knowledge modeling. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 3475–3484. (2022).
57. Zhang, H., Li, W. & Liu, Y. Comparative evaluation of knowledge embedding approaches for martial arts relationship prediction. *Knowl. Based Syst.* **258**, 110182 (2023).
58. Chen, J., Li, T. & Wang, R. Multi-level attention mechanisms for spatial-temporal feature extraction in martial arts video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7389–7398. (2022).
59. Wu, Y., Zhang, M. & Li, J. Domain-adapted transformers for martial arts textual understanding: capturing technical terminology and conceptual foundations. *Nat. Lang. Eng.* **29** (2), 245–267 (2023).
60. Yang, C., Liu, R. & Chen, L. Spatial-temporal graph convolutional networks for skeleton-based martial arts motion analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **33** (7), 3118–3132 (2022).
61. Wang, T., Zhang, H. & Chen, Y. Knowledge-guided cross-modal alignment for martial arts movement analysis: preventing spurious correlations through domain constraints. *Pattern Recognit.* **136**, 109171 (2023).
62. Li, W., Zhao, J. & Wang, L. Zero-shot and few-shot learning in cross-modal martial arts technique recognition: Leveraging knowledge graph relationships. In *Proceedings of the European Conference on Computer Vision* 289–305. (2022).
63. Zhang, M., Li, T. & Wang, Y. Transformer-based motion generation with knowledge-conditioned attention mechanisms. In *Proceedings of the International Conference on Machine Learning* 15782–15791. (2023).
64. Chen, K., Wang, Z. & Liu, J. Multi-level discrimination strategies for human motion evaluation: from joint-level assessment to holistic technique appraisal. *IEEE Trans. Hum. Mach. Syst.* **52** (3), 367–379 (2022).
65. Li, Y., Chen, J. & Zhang, L. Direct knowledge injection methodologies in deep generative frameworks: applications to martial arts motion synthesis. *Neural Netw.* **161**, 368–382 (2023).
66. Wang, R., Liu, H. & Chen, T. Graph attention networks for dynamic knowledge subgraph processing in motion generation tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence* 7842–7850. (2022).
67. Chen, Y., Zhang, W. & Li, M. Knowledge consistency constraints for physically and stylistically authentic martial arts motion generation. In *Proceedings of the IEEE International Conference on Robotics and Automation* 8754–8760. (2023).
68. Zhang, L., Wang, J. & Chen, R. Hierarchical phase modeling in traditional martial arts: decomposing complex techniques for enhanced Temporal constraint specification. *Comput. Animat. Virtual Worlds*, **33**(3–4), e2015 (2022).
69. Liu, T., Chen, K. & Wang, Y. Knowledge-enhanced generative approaches for limited-data domains: advances in cultural heritage documentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (6), 6842–6857 (2023).
70. Wang, L., Zhang, T. & Chen, Y. Multimodal martial arts dataset construction: protocols, challenges, and quality control mechanisms. *Data Brief.* **42**, 108293 (2022).
71. Chen, J., Wang, Z. & Li, T. Structured annotation methodologies for martial arts technique documentation: balancing procedural detail with conceptual depth. *J. Doc.* **79** (1), 97–115 (2023).
72. Li, W., Zhang, H. & Wang, Y. Semi-automated alignment between multimodal data and knowledge graph entities in martial arts documentation. *Inf. Process. Manag.* **59** (3), 102908 (2022).
73. Zhang, M., Chen, T. & Wang, L. Computational infrastructure for knowledge-enhanced generative modeling: hardware configurations and software ecosystems. *J. Supercomput.* **79** (5), 6123–6142 (2023).
74. Wang, Z., Li, Y. & Chen, J. Comprehensive evaluation metrics for motion reconstruction in cultural heritage applications. *Multimed. Tools Appl.* **81** (8), 11437–11456 (2022).
75. Chen, Y., Wu, Z. & Zhang, T. Optimization strategies for knowledge-enhanced generative adversarial networks in martial arts motion synthesis. *J. Real-Time Image Proc.* **20** (1), 28–43 (2023).
76. Ionescu, C., Papava, D., Olaru, V. & Sminchisescu, C. Human3.6 M: large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **36** (7), 1325–1339 (2014).
77. Wang, J., Chen, K. & Zhang, L. Knowledge-guided evaluation metrics for martial arts motion assessment: beyond kinematic accuracy. *IEEE Trans. Cybern.* **52** (11), 11784–11797 (2022).
78. Li, T., Wang, Y. & Chen, J. Knowledge-constrained regularization for complex martial arts technique reconstruction: addressing biomechanical ambiguity through domain principles. *Pattern Recogn.* **138**, 109218 (2023).
79. Chen, R., Zhang, H. & Wang, L. Comparative analysis of internal and external martial arts styles: implications for motion capture and reconstruction methodologies. *Sports Eng.* **25** (1), 12 (2022).
80. Zhang, T., Li, W. & Wang, Z. Limitations in extreme motion capture: quantifying information loss in high-velocity martial arts movements. *IEEE Sens. J.* **23** (4), 3829–3842 (2023).
81. Chen, K., Li, W. & Zhang, M. Systematic ablation study methodologies for knowledge-integrated generative models: quantifying component contributions. In *Proceedings of the International Conference on Learning Representations* 1–15 (2023).
82. Zhang, H., Wang, Y. & Chen, J. The critical role of domain-specific knowledge in martial arts motion generation: evidence from comparative performance analysis. *Neural Comput.* **34** (8), 1785–1812 (2022).
83. Li, T., Chen, R. & Wang, Z. Cross-modal attention mechanisms for bridging visual appearances and biomechanical principles in martial arts technique analysis. *IEEE Trans. Multimed.* **25** (5), 2837–2851 (2023).
84. Wang, L., Zhang, H. & Li, W. Temporal constraint mechanisms for continuous versus segmented martial arts movements: style-specific considerations. In *Proceedings of the ACM International Conference on Multimedia* 2458–2467 (2022).
85. Chen, J., Wang, Y. & Zhang, T. Parameter sensitivity quantification methodologies for complex generative models: applications in cultural heritage preservation. *J. Cult. Anal.* **8** (2), 124–143 (2023).
86. Li, M., Chen, K. & Wang, L. Optimal loss balancing strategies for knowledge-enhanced motion generation in traditional martial arts. In *Proceedings of the International Conference on Learning Representations* 1–15 (2022).
87. Wang, Y., Li, T. & Chen, J. Knowledge integration strategies for preserving subtle technical nuances in martial arts motion reconstruction. *IEEE Trans. Cult. Herit.* **15** (2), 467–483 (2023).
88. Chen, K., Zhang, H. & Wang, L. Experimental validation of knowledge-enhanced generative models across diverse martial arts styles: technical accuracy and semantic fidelity assessment. *Pattern Recogn.* **146**, 110097 (2024).
89. Li, W., Wang, Z. & Chen, Y. Domain-specific knowledge versus data-driven learning in martial arts motion generation: complementary strengths and integration approaches. *Neural Netw.* **159**, 247–262 (2023).
90. Wang, Y., Zhang, T. & Chen, J. Current limitations in martial arts motion reconstruction: rapid movements, rare techniques, and knowledge representation challenges. *Comput. Vis. Image Underst.* **229**, 103608 (2024).
91. Zhang, H., Li, T. & Wang, Y. Future directions in interactive martial arts learning systems: wearable sensing, augmented feedback, and adaptive guidance. *IEEE Trans. Hum. Mach. Syst.* **53** (1), 58–72 (2023).
92. Chen, J., Wang, Z. & Li, W. Cross-domain applications of knowledge-enhanced generative frameworks: from martial arts to diverse intangible cultural heritage preservation. *Digit. Appl. Archaeol. Cult. Herit.*, **27**, e00262 (2024).

Author contributions

Xiaoyu Yue: Conceptualization, methodology, knowledge graph design, supervision, writing - original draft,

funding acquisition. Lulu Zhang: Cross-modal feature extraction architecture, generative adversarial network implementation, experimental design, data analysis, writing - review & editing. All authors reviewed and approved the final manuscript.

Funding
Research Achievement of Jiangsu Higher Education Institutions' Education Informatization Research Association Project Number: 2025JSETKT124.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-36095-z.

Correspondence and requests for materials should be addressed to X.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.