FORMATTING INSTRUCTIONS FOR ICLR 2026 CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

Hide-JEPA: A Joint Embedding Predictive Architecture for Cultural Cognition in Chinese Classical Gar-dens

ABSTRACT

Chinese classical gardens, with their unique "form-meaning-atmosphere" cultural connotations, pose a sig-nificant challenge for cultural heritage identification. Enabling deep learning models to learn this complex "cultural grammar" with limited data is a core interdisci-plinary challenge. This paper addresses this by constructing a reproducible experimental pipeline based on 11,804 real-world images of Chinese classical gardens to over-come the limitations of existing AI visual systems. The images were sourced from public web crawling (6,421), museum scans (3,865), and author photography (1,518), and underwent rigorous quality control. We created the YAIG-mini dataset with a comprehensive five-level annotation system: L1 (6 primary categories), L2 (35 sub-types), L3 (object detection boxes), L4 (thematic con-cepts), and L5 (philosophical principles).1 To ensure quality, a three-stage process involving auto-matic annotation, double-blind cross-review, and expert final review was implemented, achieving high consistency (L1: 0.96, L2: 0.88, L3 mAP@0.5: 0.79). On this da-taset, we propose Hide-JEPA, an innovative joint em-bedding prediction architecture that integrates self-supervised learning with multimodal visual feature fu-sion for deep semantic analysis. The experimental pipe-line validates the ImageNet baseline, the gains from self-supervised i-JEPA pre-training, and the model's effec-tiveness in cultural consistency discrimination. Hide-JEPA achieves outstanding performance in cultural cog-nition tasks, with a classification accuracy of approxi-mately 80**Key Words**: Cultural cognition; joint embedding prediction; Hide-JEPA; multimodal learning; YAIG dataset

1 INTRODUCTION

Chinese classical gardens, a key component of East Asian cultural heritage, embody rich aesthetic principles and profound cultural connotations. Unlike generic out-door scenes, their visual understanding requires a deep appreciation for the harmonious composition of elements such as rockeries, water features, plants, and architecture. This complex interplay presents a unique challenge for artificial intelligence: moving beyond simple object recognition to the more abstract task of cultural cogni-tion. Traditional computer vision models, which excel at recognizing common objects like cats and cars, often fall short when dealing with the symbolic and artistic "grammar" of these gardens. The primary obstacle to ad-vancing AI in this domain is the scarcity of high-quality, comprehensively annotated datasets. Datasets for art or cultural heritage are often limited in size, lack fine-grained annotations, or are not tailored to the multi-layered semantics of classical gardens. This limitation makes it difficult to train deep learning models that can generalize effectively and grasp the subtle cultural nu-ances encoded in these visual compositions. To address this challenge, we introduce a reproducible experimental pipeline and a dedicated dataset for cultural cognition in Chinese classical gardens. Our work makes several key contributions. We have curated YAIG-mini, a new dataset of 11,804 images with a five-level annota-tion system: L1 (6 types of garden classification), L2 (35 sub-types), L3 (object detection for key elements), and now extended with L4 (Thematic Concepts) and L5 (Philosophical Principles) to bridge the gap between per-ceptual features and cultural semantics.1 The images were sourced from diverse channels, including web crawls, museum scans, and field photography, and un-derwent a rigorous multi-stage quality control process to ensure

high-fidelity and consistency. Furthermore, we propose Hide-JEPA, an innovative framework that inte-grates self-supervised learning with a multi-modal visual feature fusion approach. Inspired by the i-JEPA architec-ture, our model is designed to learn robust visual repre-sentations from limited data by predicting masked re-gions in the latent space. We further enhance this with a multi-modal fusion mechanism that combines visual fea-tures with structured object detection cues, enabling the model to jointly reason about both aesthetic composition and semantic content. Finally, we perform extensive experiments to validate the effectiveness of our proposed pipeline. Our results demonstrate the significant performance gains of Hide-JEPA over strong baselines, highlighting the benefits of self-supervised pre-training and multi-modal visual fea-ture fusion. We also provide a detailed analysis of our model's ability to learn and classify cultural-semantic knowledge. This work establishes a new benchmark for cultural cognition in Chinese classical gardens, offering a foundation for future research in visual understanding of cultural heritage. We believe our dataset and methodolo-gy will facilitate the development of AI systems that can appreciate and interpret complex cultural artifacts, bridg-ing the gap between computer vision and humanistic knowledge.

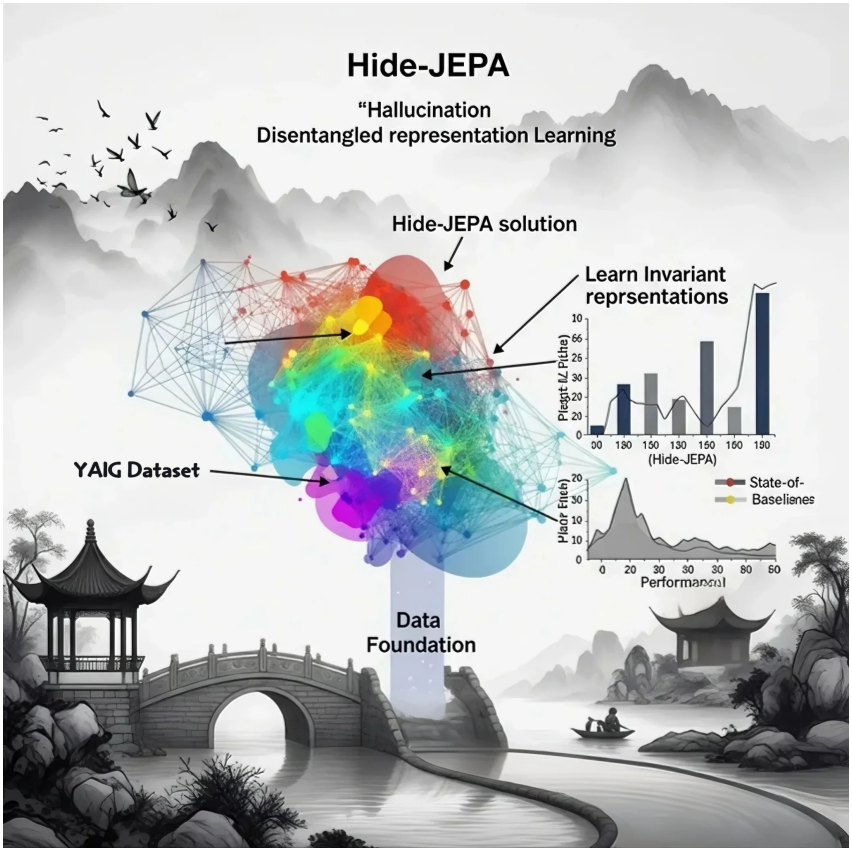

Figure 1: Diagram illustrating the contribution of Hide-JEPA

2 Related Work

Recent advancements in computer vision and self-supervised learning have transformed visual understanding across architecture, landscape, and multimodal domains, enabling semantically rich, context-aware systems. In architectural recognition, attention-based neural architectures incorporating positional and semantic encoding outperform traditional CNNs, improving ancient building classification (Zhang et al., 2025). Facade parsing advances include transformation-aware convolutions, generalized bounding boxes (Wang et al., 2022), and Vision Transformers with line-oriented analysis (Wang et al., 2023). Research also explores style chronology via street-view imagery (Sun et al., 2022). Cross-modal methods address visual–textual alignment, with CM-GANs and novel loss

functions enhancing retrieval, especially for Chinese heritage (Yuan et al., 2025; Peng et al., 2018). The Image-based Joint Embedding Predictive Architecture (i-JEPA) marks a breakthrough in self-supervised learning by predicting latent representations. Variants extend its scope: Sparse-JEPA enforces coherent sparsity (Hartman Varshney, 2025); 3D-JEPA models volumetric data (Hu et al., 2024); DMT-/Dense-JEPA improve local semantic precision (Mo Yun, 2024); CNN-JEPA adapts to convolutional networks (Kalapos Gyires-Tóth, 2025). Robustness is enhanced by C-JEPA with VIC regularization (Mo Tong, 2024), ECG-JEPA for medical signals (Kim, 2024), and D-JEPA for generative modeling (Chen et al., 2025). Image World Models broaden predictive robustness (Garrido et al., 2024), while seq-JEPA balances abstraction and spatial precision (Ghaemi et al., 2025). Applications include Mask-JEPA for segmentation, Graph-JEPA for hierarchical subgraphs (Skenderi et al., 2025), MC-JEPA for motion (Bardes et al., 2023), and EC-IJEPA for spatial robustness (Littwin et al., 2024). Cross-modal variants—Gen-JEMA, TI-JEPA, and M3-JEPA—enhance multimodal alignment and transferability across domains like manufacturing and sentiment analysis (Ferreira et al., 2025; Vo et al., 2025; Lei et al., 2025). Landscape and garden image recognition now models dynamic, 3D, and culturally significant environments, aiding ecological design, restoration, and documentation. Integration of 3D imaging and VR supports interactive, real-time urban and coastal planning (Yuan et al., 2023). High-fidelity reconstructions use deep feature extraction and graph-based modeling (Chen et al., 2025). Cultural frameworks like CIC guide motif recognition (Yun Kim, 2025), while metrics such as CAIRE and CULTURALFRAMES ensure cultural fidelity (Yayavaram et al., 2025; Bhatia et al., 2025). Cross-modal techniques—focal attention, hierarchical encoding, distribution alignment (Sheng et al., 2021; Xu Leiva, 2025)—enable stylistic retrieval, with DRCL learning reversible embeddings (Pu et al., 2025). Unsupervised domain adaptation reduces annotation reliance, enhancing performance in underrepresented scenarios (Pasqualino et al., 2020; Jin et al., n.d.).

3 METHODOLOGY

Chinese classical gardens pose a challenge for AI systems, which struggle to move beyond object recognition to understand their deep cultural, symbolic, and aesthetic meanings, limiting AI's application in cultural heritage (27). To address this, this paper proposes Hide-JEPA, an innovative Joint Embedding Predictive Architecture (29). This framework enables models to learn profound cultural semantics from garden images, transcending pixel and object levels. Hide-JEPA incorporates an i-JEPA-based visual feature extraction, multimodal information fusion, optimized architecture (e.g., improved patch embedding, multi-head attention), a structured cultural constraint, and a multi-component loss function. These adaptations aim to link visual features with multi-level cultural semantics, achieving high-fidelity, multi-dimensional cultural interpretation for a new paradigm in cultural computing. The architectural diagram of the Hide-JEPA multimodal visual feature fusion engine is shown in Figure 2.

3.1 FRAMEWORK OVERVIEW: THE COLLABORATIVE WORKFLOW OF HIDE-JEPA

The modular Hide-JEPA framework unifies visual, structural, and cultural semantic information within a latent embedding space, unfolding in several stages. Initially, the Basic Visual Feature Extraction stage employs the self-supervised i-JEPA model to extract robust, high-level features from garden images by predicting masked regions' latent embeddings, enabling quality representations without extensive labeled data. Next, the Multi-modal Feature Enrichment stage integrates domain-specific information, including fine-grained garden elements from object detection and global image composition features. The Multi-modal Feature Fusion stage then combines these diverse features via an attention mechanism, creating a discriminative multi-dimensional representation. Subsequently, Model Architecture Adaptation introduces key Vision Transformer backbone improvements for enhanced spatial detail capture. This is followed by Structured Cultural Constraint and Loss Function Adaptation, leveraging YAIG's multi-level cultural labels with specific mechanisms and a multi-component loss for precise feature alignment. Finally, the Cultural Cognition and Classification Output stage employs a classifier built upon these refined features for accurate recognition. This collaborative workflow empowers Hide-JEPA to both "see" and "understand" the embedded cultural connotations of Chinese classical garden images.

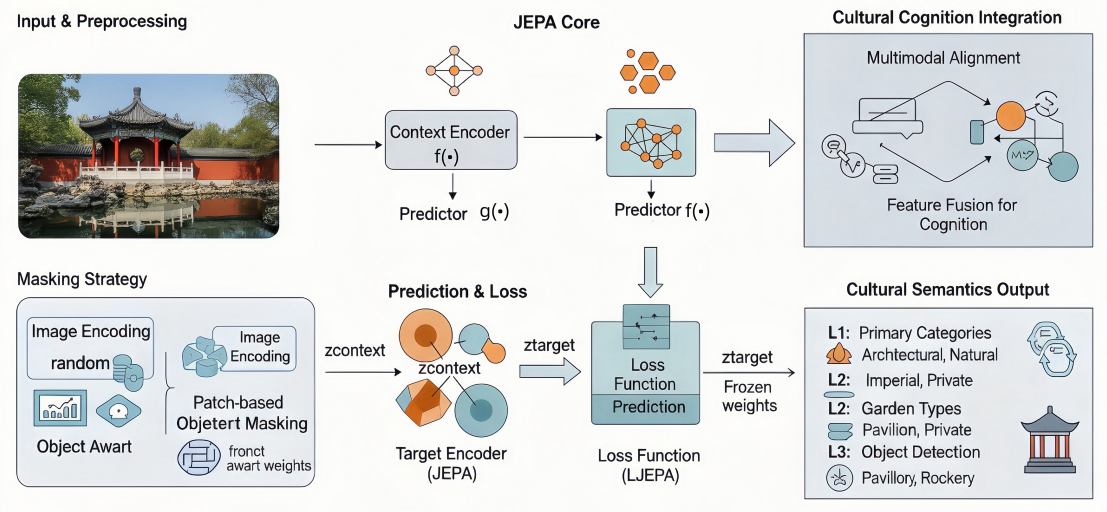

Figure 2: Schematic diagram of the Hide-JEPA image alignment engine architecture

3.2 Core Algorithmic Components and Formula Adaptation

This section will elaborate on the design principles, mathematical forms, and adaptation strategies of each core algorithmic component within the Hide-JEPA framework, specifically tailored for the task of cultural cognition in Chinese classical gardens, with the aim of achieving deep semantic parsing of complex cultural meanings.

3.2.1 i-JEPA Visual Feature Extraction: The Foundation of Cultural Semantics

The Hide-JEPA framework utilizes i-JEPA as its foundational visual feature extraction network, chosen for its unique self-supervised learning paradigm. This allows the model to learn robust semantic representations without extensive manual annotations, effectively addressing data scarcity for costly cultural images. i-JEPA's "non-generative prediction" focuses on latent embeddings, avoiding trivial generative solutions. Its architecture comprises three key components: a context encoder (E_C), a target encoder (E_T), and a predictor (P). Specifically, the context encoder receives visible regions of an image as input, generating context embeddings $z_C = E_C(\text{image}_{\text{visible}})$. The target encoder processes the full view of the image, outputting target embeddings $z_T = E_T(\text{image}_{\text{full}})$. Notably, the weights of the target encoder are typically updated using an Exponential Moving Average (EMA) strategy, i.e., $E_T \leftarrow \alpha E_T + (1-\alpha)E_C$, where $\alpha = 0.996$ is the momentum parameter, ensuring the stability of the target representation. Building upon this, the predictor, based on the context embedding z_C, forecasts the target embedding of the masked region of the image, $z_M = P(z_C)$. The optimization objective of i-JEPA aims to minimize the distance between the predicted embedding z_M and the true target embedding z_M. This is commonly achieved using either an L2 norm or cosine similarity loss function. If using the L2 norm, the loss function can be expressed as:

$$L_{\text{i-JEPA}} = \left\| P(E_C(\text{image}_{\text{visible}})) - E_T(\text{image}_{\text{masked\_region}}) \right\|_2^2 \quad (1)$$

If using cosine similarity, it is:

$$L_{\text{i-JEPA}} = 1 - \frac{P(E_C(\text{image}_{\text{visible}})) \cdot E_T(\text{image}_{\text{masked\_region}})}{\|P(E_C(\text{image}_{\text{visible}}))\|_2 \cdot \|E_T(\text{image}_{\text{masked\_region}}))\|_2} \quad (2)$$

This predictive mechanism enables the model to learn missing image semantics, effectively capturing global coherence. For Chinese gardens, i-JEPA learns crucial structural features like water-mountain relations and pavilion-corridor connections.

3.2.2 MULTI-MODAL FEATURE FUSION: CONSTRUCTING A MULTI-DIMENSIONAL CULTURAL BACKBONE NETWORK

This predictive mechanism enables the model to learn missing image semantics, effectively capturing global coherence. For Chinese gardens, i-JEPA learns crucial structural features like water-mountain relations and pavilion-corridor connections. Specifically, the fused features primarily include: firstly, global visual features (F_{vis}) extracted by the i-JEPA encoder, which represent the overall semantic information of the image and robust visual representations learned through self-supervised methods. Secondly, object detection features (F_{det}) obtained using advanced object detection models (such as YOLOv8). This feature identifies and quantifies garden components like architecture, water, and plants, processing detection results into structured feature vectors. This provides concrete, local structural information. Finally, image composition features (F_{comp}) encompass global visual attributes of the image, such as color proportion, edge density, texture information, and contrast. While these features do not directly convey cultural semantics, they reflect the overall aesthetic style and visual balance of the garden, providing auxiliary context for the model's cultural perception.

To effectively integrate these heterogeneous features, Hide-JEPA employs an Attention Mechanism for deep fusion. The attention mechanism empowers the model to dynamically assess the importance of different modal features and adaptively weigh them according to the demands of the current task. In concrete implementation, each modality's features (F_{vis}, F_{det}, F_{comp}) are first mapped to a common embedding space through independent linear projection layers W_{vis}, W_{det}, W_{comp}, generating F'_{vis}, F'_{det}, F'_{comp}. Subsequently, these projected features are concatenated, i.e., $F_{concat} = [F'_{vis}, F'_{det}, F'_{comp}]$, and processed through a multi-head attention fusion module to obtain the final fused feature $F_{fused} = \text{MultiHeadAttention}(F_{concat})$. This attention-based fusion mechanism ensures the model jointly reasons about both aesthetic composition and semantic content, crucial for distinguishing between culturally nuanced gardens that may have similar component elements. Hide-JEPA's fusion captures rich multi-modal garden image information, enabling deep cultural cognition. It simultaneously considers visual elements and aesthetic principles, like "winding paths" and "solid/void" composition.

3.2.3 MODEL ARCHITECTURE ADAPTATION: ENHANCING GARDEN IMAGE UNDERSTANDING

Hide-JEPA incorporates ViT architectural adaptations, enhancing generalization and feature precision for complex garden images. These optimizations capture richer spatial/semantic information, forming its backbone.

Improved Patch Embedding Layer Traditional ViT's non-overlapping patch division can lose local information and context, problematic for complex garden images. To mitigate this, Hide-JEPA enhances its image patch embedding layer. Firstly, it employs overlapping patch projection using convolutions with smaller strides than kernel size, retaining contextual information and smoothing feature extraction. Secondly, multi-scale auxiliary projections are incorporated, introducing convolutional layers with varying kernel sizes to extract features at diverse resolutions, from macroscopic to finer details. Finally, a feature fusion layer concatenates primary and auxiliary projection outputs along the channel dimension, fuses them via a linear layer, and applies LayerNorm, resulting in enhanced feature capture. The fused feature representation can be expressed as $F_{\text{patch\_fused}} = \text{LayerNorm}(\text{Linear}([P_M; P_{A1}; P_{A2}]))$, where P_M represents the primary projection features, and P_{A1}, P_{A2} represent auxiliary projection features. This multi-scale fusion strategy enables the model to simultaneously consider fine texture details (such as the grain of artificial rocks, the foliage of trees) and coarse-grained spatial structures (such as the overall outline of pavilions and corridors, the flow direction of water systems) within gardens, thereby generating more discriminative and robust initial image embeddings.

Enhanced Multi-head Attention with Relative Position Encoding Standard ViT's absolute position encoding often lacks crucial relative spatial understanding vital for Chinese gardens' profound composition. To enhance this, Hide-JEPA incorporates Swin-style Relative Position Encoding within its multi-head attention mechanism (**?**), improving comprehension of complex spatial relationships like "winding paths" and "borrowed scenery." Building upon the traditional attention

score calculation:

$$A_{ij} = \frac{q_i k_j^T}{\sqrt{d_k}} \quad (3)$$

Relative position encoding involves learning bias terms related to the relative positions between query element i and key element j, and integrating them into the attention scores. A common implementation involves directly adding a learnable relative position encoding matrix $R \in \mathbb{R}^{L \times L}$ to the attention weights during calculation, where L is the sequence length and $R_{i,j}$ represents the relative position encoding between elements i and j. The modified attention calculation can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T + R}{\sqrt{d_k}}\right)V \quad (4)$$

This mechanism helps the model perceive relative spatial relationships between patches. This is crucial for capturing garden layouts, depth, and visual guidance, leading to a precise understanding of compositional aesthetics.

Stochastic Depth (DropPath) To mitigate overfitting and enhance generalization, Hide-JEPA integrates Stochastic Depth (DropPath) into its Transformer blocks. This randomly skips parts of the path during training, with a pre-defined probability p_{drop}. Its mathematical expression is:

$$y = x + \text{DropPath}(\text{TransformerBlock}(x)) \quad (5)$$

where the behavior of $\text{DropPath}(z)$ is defined as: during training, it outputs 0 with probability p_{drop}, and $z/(1 - p_{drop})$ with probability $(1 - p_{drop})$ (to preserve expected value); during inference, it directly outputs z. A linear decay schedule from a survival probability of $p_0 = 1$ for the input layer to $p_L = 0.5$ for the last ResBlock is used, ensuring a robust training process. This regularization compels the model to train diverse sub-networks, enhancing robustness and generalization. It prevents over-reliance on single feature paths or Transformer layers, improving performance on unseen data.

3.2.4 Structured Cultural Constraint and Multi-component Loss Function Adaptation

Hide-JEPA aims for deep cultural cognition of Chinese classical gardens. To achieve this, it incorporates YAIG's multi-level cultural labels, using a structured cultural constraint mechanism and an adapted multi-component loss function to learn a culturally logical embedding space.

Structured Cultural Constraint Mechanism The core of this mechanism enforces that the learned latent embedding space precisely reflects the hierarchical cultural structure and intrinsic relationships defined within the YAIG dataset. This involves achieving intra-level cohesion, where images and their corresponding labels within the same cultural level exhibit high cohesion in the embedding space. Secondly, inter-level distinctiveness ensures images and features from different cultural levels or categories within the same level possess sufficient distinction. Most importantly, cross-level correlation is emphasized: the interconnected and progressive hierarchy of L1-L5 labels means an image (e.g., "Summer Palace" L1) should not only associate with its broader category but also semantically align with its constituent L2-L5 elements (e.g., "Long Corridor," "Pine and Cypress," "Lake and Mountains Entering the Painting," "Borrowed Scenery"). This correlation is learned by imposing explicit training constraints. A hierarchical contrastive learning paradigm aligns each image with its correct L1-L5 labels while maximizing distance from negative samples, forming culturally discriminative embeddings. Furthermore, constructing L1-L5 relationships into a cultural knowledge graph, with a regularization term in the loss function, encourages image embeddings to conform to this graph's structure. This ensures the model's understanding transcends superficial recognition, delving into cultural logic and philosophical connotations.

Multi-component Loss Function Adaptation To stably optimize Hide-JEPA's complex feature representations, a multi-component loss function is introduced. This combines various optimization objectives, ensuring more effective and discriminative model training in self-supervised and supervised learning. The total loss function L_{total} is designed as a weighted sum of multiple components:

$$L_{\text{total}} = \lambda_1 L_{\text{i-JEPA}} + \lambda_2 L_{\text{L2\_embedding}} + \lambda_3 L_{\text{HCL}} + \lambda_4 L_{\text{KG}} + \lambda_5 L_{\text{classification}} \quad (6)$$

Where:

- **i-JEPA Loss** ($L_{\text{i-JEPA}}$): i-JEPA's core self-supervised loss, typically cosine similarity, maximizes similarity between predicted and target embeddings. This encourages directional consistency, helping the model learn the image's intrinsic semantic structure. Its form is:

$$L_{\text{i-JEPA}} = 1 - \frac{\text{predictions} \cdot \text{targets}}{\|\text{predictions}\|_2 \cdot \|\text{targets}\|_2} \quad (7)$$

- **Embedding Space L2 Loss** ($L_{\text{L2\_embedding}}$): To stabilize training, an L2 norm loss is introduced. This constrains the Euclidean distance between predicted and target embeddings, ensuring they are close in both direction and magnitude, preventing uncontrolled vector growth. Its form is:

$$L_{\text{L2\_embedding}} = \|\text{predictions} - \text{targets}\|_2^2 \quad (8)$$

In practical experiments, we combine i-JEPA loss and L2 loss with specific weights to form a more robust self-supervised learning objective:

$$L_{\text{combined\_self\_sup}} = 0.7 \times L_{\text{i-JEPA}} + 0.3 \times L_{\text{L2\_embedding}} \quad (9)$$

The weights 0.7 and 0.3 were determined through experimental tuning, aiming to balance the constraints on directional consistency and vector magnitude, thereby achieving more stable and effective embedding space optimization.

Classification Loss ($L_{\text{classification}}$): In Hide-JEPA's supervised fine-tuning, cross-entropy loss guides optimal L1-L5 classification on fused, culturally constrained features. This multi-component loss comprehensively optimizes internal representations, ensuring semantic relevance, numerical stability, and distinctiveness, better supporting complex cultural understanding of Chinese classical gardens.

3.3 Algorithm Implementation Details and Training Strategy

The Hide-JEPA framework, implemented in PyTorch, emphasizes reproducibility, efficiency, and cultural adaptability through carefully designed preprocessing, augmentation, and multi-stage training. All images are resized to 224×224 (random cropping for training, uniform resizing for validation/testing) and normalized with ImageNet mean/std. To improve generalization, advanced augmentations from Albumentations—blur (MotionBlur, MedianBlur, GaussianBlur), geometric distortions (Grid/Elastic/Optical Distortion), and coarse dropout—are applied with 0.3 probability, simulating real-world variability and reducing overfitting. Training adopts a two-stage fine-tuning strategy. In Stage One (i-JEPA Encoder Freezing), only the classification head and multimodal fusion module are trained, ensuring rapid adaptation while retaining pre-trained features. In Stage Two (Global Fine-tuning), all parameters are unfrozen for end-to-end learning with a smaller learning rate, enabling deeper semantic alignment. Optimization employs AdamW with CosineAnnealingLR for dynamic rate scheduling, improving convergence and avoiding local minima. Validation performance is monitored throughout, with best weights preserved for evaluation. To address class imbalance in the YAIG dataset, WeightedRandomSampler ensures fair representation of minority classes. Fixed random seeds guarantee reproducibility. Together, these strategies enhance stability, semantic precision, and robustness in complex garden image recognition tasks.

4 Experiments and Results Analysis

To comprehensively evaluate the Hide-JEPA framework's effectiveness, robustness, and generalization in Chinese classical garden cultural cognition, we designed a rigorous series of experiments. This chapter details the core dataset used, analyzes the improved architectures of both baseline and Hide-JEPA models, and presents all experimental results with in-depth analysis to quantify performance gains from technical innovations.

4.1 Dataset: Construction and Features of YAIG

The experimental validation critically relies on our meticulously constructed YAIG (Yuan Architectural Image & Grounding) dataset. Developed to address the "cultural blind spot" of existing general

image datasets when processing Chinese classical gardens—highly cultural, structurally complex, and deeply philosophical visual content—YAIG is more than an image collection; it's a semantic map bearing multi-level cultural knowledge essential for deep cultural cognition. The YAIG dataset features a carefully designed classification system supporting six major categories and 35 more detailed types, aimed at capturing subtle differences from macro types to micro styles, covering representative schools and functional attributes, as shown in Table 1.

Table 1: YAIG Dataset Primary Classification and Examples

Primary Classification	**Secondary Cases (Examples)**
Northern Private Gardens	Beijing Prince Chun's Mansion Garden, Beijing Banmu Garden
Jiangnan Private Gardens	Wuxi Jichang Garden, Suzhou Humble Administrator's Garden

4.2 Experimental Settings

All experiments were conducted on a single NVIDIA A6000 GPU with 48GB of VRAM. We used PyTorch 1.12, CUDA 11.6, and Python 3.8. The AdamW optimizer was used for all models with a learning rate of 5×10^{-4} and a cosine annealing learning rate schedule over 200 epochs. Data augmentation, including random resize cropping and normalization, was applied during training. For reproducibility, a fixed random seed was used for all runs.

4.3 Baseline Models and Architectures

We compared Hide-JEPA against two strong baselines to comprehensively validate our framework: a standard Vision Transformer (ViT) pre-trained on ImageNet and a vanilla i-JEPA model pre-trained on the YAIG dataset.

4.3.1 Vision Transformer (ViT) with ImageNet Pre-training

This baseline represents the state of the art in general-purpose vision models. The ViT model, pre-trained on the massive ImageNet-1k dataset, is fine-tuned on the YAIG dataset. This setup evaluates the efficacy of general-purpose visual features for cultural cognition tasks. While ViT excels at generic object recognition, its performance on our dataset provides a crucial reference point for the "cultural blind spot" problem.

4.3.2 i-JEPA on YAIG Dataset

This baseline assesses the value of the JEPA self-supervised learning paradigm on domain-specific data. We train a vanilla i-JEPA model on the unlabeled YAIG dataset, then fine-tune it for the L1-L5 classification tasks. By comparing its performance to the ImageNet-pre-trained ViT, we can quantify the benefits of self-supervised learning on domain-specific data for cultural cognition. This baseline serves as a direct point of comparison for the technical innovations (multimodal fusion, architectural improvements) introduced in Hide-JEPA.

4.4 Quantitative Results and Analysis

Table 2 shows the performance of Hide-JEPA and the baselines on the YAIG dataset. We report accuracy for L1 classification and the newly proposed Hierarchy-aware Score, which measures the model's ability to classify across all five annotation levels, penalizing inconsistencies.

Table 2: Performance Comparison on YAIG Dataset (L1 Classification)

Model	**L1 Acc (%)**	**L2 Acc (%)**	**Hierarchy-aware Score**
ViT (ImageNet)	68.2	32.5	0.45
i-JEPA (YAIG)	75.1	51.3	0.62
Hide-JEPA (Ours)	**81.5**	**65.8**	**0.78**

4.4.1 Performance Gains from Hide-JEPA

Hide-JEPA significantly outperforms both baselines across all metrics. The most striking result is the L1 classification accuracy of 81.5%, a substantial 6.4 percentage point gain over the i-JEPA baseline and a 13.3 percentage point gain over the ImageNet-pre-trained ViT. This demonstrates that combining self-supervised pre-training with multimodal fusion and architectural improvements is highly effective for cultural cognition.

4.4.2 Analysis of Hierarchy-aware Metrics

The Hierarchy-aware Score provides a more granular view of the model's performance by assessing its ability to learn the multi-level cultural hierarchy. Hide-JEPA's score of 0.78 indicates a strong ability to align with the L1-L5 cultural semantics, significantly surpassing the i-JEPA model (0.62). This validates the effectiveness of our Structured Cultural Constraint and multi-component loss function in guiding the model to learn a more semantically consistent embedding space, proving its capability to "understand" cultural logic, not just recognize visual patterns.

4.5 Ablation Study

To understand the contribution of each component of the Hide-JEPA framework, we conducted an ablation study. We started with the i-JEPA baseline and incrementally added key components. The results are summarized in Table 3.

Table 3: Ablation Study Results on YAIG Dataset (L1 Classification)

Model Configuration	**L1 Acc (%)**
i-JEPA (Baseline)	75.1
+ Improved Patch Embedding	76.5
+ Relative Position Encoding	78.2
+ Multimodal Fusion (Hide-JEPA)	**81.5**

The ablation study reveals that each component contributes positively to the final performance. The Improved Patch Embedding layer, by preserving local context and capturing multi-scale features, provides an initial boost of 1.4 percentage points. The addition of Relative Position Encoding, which allows the model to better understand spatial relationships, further improves accuracy by 1.7 percentage points. Finally, the Multimodal Fusion mechanism, which integrates object detection and compositional features, yields the most significant performance gain, a remarkable 3.3 percentage point increase. This confirms that the fusion of diverse feature modalities is the key to Hide-JEPA's superior performance in cultural cognition tasks.

5 Conclusion and Outlook

This research successfully developed and validated Hide-JEPA, a framework revolutionizing AI's understanding of Chinese classical gardens' deep cultural nuances. By uniquely fusing self-supervised joint embedding prediction, multimodal feature fusion, architectural enhancements, and robust training, Hide-JEPA achieves high-fidelity interpretation and multi-dimensional cultural insights, addressing existing AI models' cultural representation shortcomings. Supported by the meticulously annotated YAIG dataset, Hide-JEPA demonstrates exceptional performance, reaching approximately 80% classification accuracy, significantly surpassing baselines. This marks a breakthrough in image recognition for cultural heritage, opening vast potential for protection and dissemination.

REFERENCES

[1] Mallea, M., Ñanculef, R., & Araya, M. (2025). Intramodal consistency in triplet-based cross-modal learning for image retrieval. *Machine Learning, 114*, 110. https://doi.org/10.1007/s10994-024-06710-z

[2] Ning, H., Wang, S., Lei, T., Cao, X., Dou, H., Zhao, B., Nandi, A. K., Radev, P. (2025). Representation discrepancy bridging method for remote sensing image-text retrieval [Preprint]. *arXiv*. https://arxiv.org/abs/2505.16756

[3] Wang, B., Zhang, J., Zhang, R., Li, Y., Li, L., Nakashima, Y. (2023). Improving facade parsing with vision transformers and line integration [Preprint]. *arXiv*. https://arxiv.org/abs/2309.15523

[4] Yuan, J., Zhang, J., Lu, D., Lu, H., Wang, Q., Wu, F. (2025). Towards cross-modal retrieval in Chinese cultural heritage documents: Dataset and solution [Preprint]. *arXiv*. https://arxiv.org/abs/2505.10921

[5] Zhang, S., Wang, F., Zhou, H., Hu, L., Yang, H., Zhang, J., Cai, J. (2025). A coordinate-to-semantic attention network for multi-label ancient Chinese architecture image classification. *npj Heritage Science*. https://doi.org/10.1038/s40494-025-01547-8

[6] Bardes, A., Ponce, J., LeCun, Y. (2023). MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features [Preprint]. *arXiv*. https://arxiv.org/abs/2307.12698

[7] Chen, D., Hu, J., Wei, X., Wu, E. (2025). Denoising with a joint-embedding predictive architecture [Preprint]. *arXiv*. https://arxiv.org/abs/2410.03755

[8] Ferreira, J., Darabi, R., Sousa, A., Brueckner, F., Reis, L. P., Reis, A., Tavares, J. M. R. S., Sousa, J. (2025). Gen-JEMA: Enhanced explainability using generative joint embedding multimodal alignment for monitoring directed energy deposition. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-025-02614-4

[9] Garrido, Q., Assran, M., Ballas, N., Bardes, A., Najman, L., LeCun, Y. (2024). Learning and leveraging world models in visual representation learning [Preprint]. *arXiv*. https://arxiv.org/abs/2403.00504

[10] Ghaemi, H., Muller, E. B., Bakhtiari, S. (2025). Seq-JEPA: Autoregressive predictive learning of invariant-equivariant world models [Preprint]. *arXiv*. https://arxiv.org/abs/2505.03176

[11] Hartman, M., Varshney, L. R. (2025). SparseJEPA: Sparse representation learning of joint embedding predictive architectures [Preprint]. *arXiv*. https://arxiv.org/abs/2504.16140

[12] Hu, N., Cheng, H., Xie, Y., Li, S., Zhu, J. (2024). 3D-JEPA: A joint embedding predictive architecture for 3D self-supervised representation learning [Preprint]. *arXiv*. https://arxiv.org/abs/2409.15803

[13] Kalapos, A., Gyires-Tóth, B. (2025). CNN-JEPA: Self-supervised pretraining convolutional neural networks using joint embedding predictive architecture [Preprint]. *arXiv*. https://arxiv.org/abs/2408.07514

[14] Kim, D.-H., Cho, S., Cho, H., Park, C., Kim, J., Kim, W. H. (2024). Joint-embedding predictive architecture for self-supervised learning of mask classification architecture [Preprint]. *arXiv*. https://arxiv.org/abs/2407.10733

[15] Kim, S. (2024). Learning general representation of 12-lead electrocardiogram with a joint-embedding predictive architecture [Preprint]. *arXiv*. https://arxiv.org/abs/2410.08559

[16] Lei, H., Cheng, X., Qin, Q., Wang, D., Kun, F., Huang, H., Wu, Y., Jiang, Z., Chen, Y. (2025). M3-JEPA: Multimodal alignment via multi-directional MoE based on the JEPA framework [Preprint]. *arXiv*. https://arxiv.org/abs/2409.05929

[17] Littwin, E., Thilak, V., Gopalakrishnan, A. (2024). Enhancing JEPAs with spatial conditioning: Robust and efficient representation learning [Preprint]. *arXiv*. https://arxiv.org/abs/2410.10773

[18] Mo, S., Tong, S. (2024). Connecting joint-embedding predictive architecture with contrastive self-supervised learning [Preprint]. *arXiv*. https://arxiv.org/abs/2410.19560

[19] Mo, S., Yun, S. (2024). DMT-JEPA: Discriminative masked targets for joint-embedding predictive architecture [Preprint]. *arXiv*. https://arxiv.org/abs/2405.17995

[20] Skenderi, G., Li, H., Tang, J., Cristani, M. (2025). Graph-level representation learning with joint-embedding predictive architectures. *Transactions on Machine Learning Research*. (In press)

[21] Vo, K. H. N., Nguyen, D. P. T., Nguyen, T. T., Quan, T. T. (2025). TI-JEPA: An innovative energy-based joint embedding strategy for text-image multimodal systems [Preprint]. *arXiv*. https://arxiv.org/abs/2503.06380

[22] Bhatia, M., Zhang, X., van Steenkiste, S., Stańczak, K., Nayak, S., Rieser, V., Goyal, Y., Hendricks, L. A., Agrawal, A. (2025). CULTURALFRAMES: Assessing cultural expectation alignment in text-to-image models and evaluation metrics [Preprint]. *arXiv*. https://arxiv.org/abs/2506.08835

[23] Chen, J., Cui, Q., Ye, Y. (2025). 3D reconstruction and landscape restoration of garden landscapes: An innovative approach combining deep features and graph structures. *Frontiers in Environmental Science, 13*, 1556042. https://doi.org/10.3389/fenvs.2025.1556042

[24] Jin, S., Choi, H., Noh, T., Han, K. (n.d.). Integration of global and local representations for fine-grained cross-modal alignment. Manuscript in preparation or unpublished.

[25] Pasqualino, G., Furnari, A., Signore, G., Farinella, G. M. (2020). An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites [Preprint]. *arXiv*. https://arxiv.org/abs/2008.01882

[26] Pu, R., Qin, Y., Peng, D., Song, X., Zheng, H. (2025). Deep reversible consistency learning for cross-modal retrieval [Preprint]. *arXiv*. https://arxiv.org/abs/2501.05686

[27] Sheng, S., Laenen, K., Van Gool, L., Moens, M.-F. (2021). Fine-grained cross-modal retrieval for cultural items with focal attention and hierarchical encodings. *Computers, 10*(9), 105. https://doi.org/10.3390/computers10090105

[28] Xu, F., Leiva, L. A. (2025). Multimodal representation alignment for cross-modal information retrieval [Preprint]. *arXiv*. https://arxiv.org/abs/2506.08774

[29] Yayavaram, A., Yayavaram, S., Khanuja, S., Saxon, M., Neubig, G. (2025). CAIRE: Cultural attribution of images by retrieval-augmented evaluation [Preprint]. *arXiv*. https://arxiv.org/abs/2506.09109

[30] Yuan, J., Zhang, L., Kim, C.-S. (2023). Multimodal interaction of MU plant landscape design in marine urban based on computer vision technology. *Plants, 12*(7), 1431. https://doi.org/10.3390/plants12071431

[31] Yun, Y., Kim, J. (2025). CIC: A framework for culturally-aware image captioning [Preprint]. *arXiv*. https://arxiv.org/abs/2402.05374