EGF-CHITR: An Entity-Guided Framework for Cultural Heritage Cross-Modal Image–Text Retrieval

Gang Xiao[1], Yiqiang He[1], Fengyi Ye[2], Jinhui Yu[3], and Jun Xu[4,*]
[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
[2]Hangzhou Tianyuan Textile Co., Ltd, Hangzhou, China
[3]Zhejiang Provincial Museum, Hangzhou, China
[4]School of Design and Architecture, Zhejiang University of Technology, Hangzhou, China
*Corresponding author: xujun@zjut.edu.cn

***Abstract*—Chinese cultural heritage artifacts are rich in multimodal signals, making image–text cross-modal retrieval crucial for digital preservation and reuse. Yet reliable image–text alignment remains challenging due to (i) background-heavy professional descriptions that introduce redundant semantics and (ii) CLS-level global matching that can obscure locally discriminative visual evidence. We propose EGF-CHITR, a lightweight framework that integrates entity-prior gating with bidirectional local alignment to enhance cross-modal discriminability. Specifically, entity-prior gating injects domain knowledge to reweight textual tokens, while bidirectional local alignment models hierarchical patch–token interactions to suppress redundant semantics and strengthen locally discriminative correspondence under global semantic consistency. Experiments on the ceramics dataset PorTi and the public cultural-heritage dataset CulTi demonstrate that EGF-CHITR consistently outperforms mainstream methods across multiple metrics in both retrieval directions.**

***Index Terms*—Cultural Heritage Digital Preservation, Image–Text Retrieval, Entity-Prior Gating, Bidirectional Local Alignment**

I. Introduction

Digital preservation and utilization of cultural heritage are of great significance [1]. With the continuous growth of digital resources such as artifact images, collection archives, and scholarly texts [2], cross-modal image–text retrieval has become a fundamental technique to support expert-assisted research, automated cataloging, similar-object comparison, and knowledge dissemination [3]. Unlike generic visual domains, cultural-heritage objects typically exhibit pronounced intra-class similarity [4]: many artifacts look highly alike at the global appearance level, while truly discriminative evidence is often distributed in subtle and localized visual differences [5], which poses higher requirements for fine-grained understanding and robust cross-modal alignment.

Despite rapid progress in image–text retrieval, two critical challenges remain in cultural-heritage scenarios. First, image–text semantic asymmetry is prominent. Professional catalog descriptions often contain generic narratives such as craft history and period background [6], while truly discriminative fine-grained attributes related to object differentiation are relatively sparse. This makes general-purpose vision–language pre-trained models susceptible to redundant semantics and noisy supervision, resulting in biased cross-modal alignment [7]. Second, local detail alignment is difficult. Differences among cultural-heritage objects are typically entity-level, local, and structurally fine-grained [8]. However, many mainstream retrieval methods still center on global embedding matching [9] and lack explicit modeling of correspondences between local visual cues and fine-grained textual semantics [10]. Consequently, they often fail to capture key visual evidence that is low-saliency but highly discriminative, limiting both fine-grained retrieval performance and interpretability.

To address these challenges, we propose EGF-CHITR, an entity-guided fine-grained image–text retrieval framework for cultural heritage. It encodes domain entities as token-level priors to highlight visually discriminative attributes and suppress background-heavy narratives, while using proposal-free bidirectional patch–token interaction to better capture subtle, texture-dominant local evidence without relying on object proposals.

Our main contributions are as follows:

- We propose EGF-CHITR, an end-to-end, entity-prior-driven fine-grained retrieval framework tailored to the semantic asymmetry and local alignment challenges in cultural-heritage image–text retrieval.
- We design lightweight enhancement modules that integrate entity-prior gating with bidirectional local alignment, jointly strengthening the modeling and alignment of key local cues via knowledge injection and bidirectional patch–token interactions.
- We construct PorTi, an entity-level annotated dataset for the ceramics domain, and further validate the effectiveness of EGF-CHITR for cross-modal semantic association learning on CulTi.

II. Related Work

The goal of image–text retrieval is to learn a shared embedding space for semantic alignment across visual and textual modalities. Early global matching methods such as VSE++ [9] focus on coarse-grained consistency by encoding

the whole image and sentence into global vectors, but this strategy is often insufficient for cultural-heritage retrieval, where artifacts usually exhibit high intra-class similarity and differ mainly in subtle local cues [11]. To improve fine-grained sensitivity, local alignment methods such as SCAN [12] model region–word correspondences via cross-attention, typically relying on bottom-up region features produced by detection pipelines such as Faster R-CNN [13]. However, in cultural-heritage imagery, discriminative evidence is often weakly localized, texture-dominant, and not well matched to object-centric proposals, which limits the robustness of proposal-based matching. Related studies in the artistic domain have also explored semantic understanding and multimodal retrieval, further showing the potential of vision–language techniques for cultural-content analysis [14]. In parallel, vision–language pre-training methods represented by CLIP [15] and Chinese-CLIP [16] have established strong global alignment capabilities, while prompt learning and parameter-efficient adaptation methods, including CoOp [17], CoCoOp [18], adapters [19], [20], and LoRA [21], improve adaptation efficiency. Nevertheless, these methods mainly recalibrate global representations and do not explicitly inject domain entity structure into token-level semantic modeling, which is problematic for cultural-heritage descriptions containing redundant narratives and sparse discriminative attributes.

Recent research has moved toward deeper cross-modal interaction and fine-grained retrieval. ALBEF [22] strengthens multimodal fusion through cross-modal attention, while knowledge-enhanced approaches such as ERNIE-ViL [23] introduce structured knowledge into vision–language representation learning. Dataset construction has also been explored in the artistic domain; for example, Artpedia provides paired visual and contextual descriptions for artworks, offering a useful benchmark for visual–semantic learning beyond generic image–text datasets [24]. Fine-grained retrieval methods further improve local discrimination through more sensitive interaction mechanisms, such as FLMR [25] for late interaction and MP-FGVC [26] for multimodal prompting; RegionCLIP [27] also explores region-level alignment under proposal-based settings. Despite these advances, existing methods still face three limitations in fine-grained cultural-heritage retrieval: they often lack explicit token-level entity priors, still depend on salient regions or proposal quality, and rarely impose bidirectional consistency constraints on local cross-modal focus. To address these issues, our framework incorporates domain entity spans as structured token-level priors, performs proposal-free bidirectional patch–token interaction, and regularizes the two interaction directions with an attention-consistency term, making it better suited to cultural-heritage scenarios in which discriminative evidence is subtle, texture-dominant, and weakly localized.

III. Method

As illustrated in Fig. 1, EGF-CHITR adopts the pre-trained Chinese vision–language model CN-CLIP as the backbone for cross-modal representation learning. In Fig. 1, H_I and H_T denote the original image and text token sequences, $H'T$ the entity-reweighted text sequence, r the binary entity-indicator sequence, uI and vT the final normalized embeddings, and Lcc, $LI \to T$, and $LT \to I$ the training losses. The framework combines entity-prior gating for text reweighting with bidirectional local alignment for proposal-free patch–token interaction, improving both core-attribute modeling and fine-grained correspondence.

A. Backbone Model and Representations

We adopt ViT-B/16 as the visual encoder and RoBERTa as the text encoder, both initialized from publicly available pre-trained weights. Given an input image I and its Chinese description T, the encoders output token-sequence representations for the image and text, respectively:

$$H_I = \{p_0, p_1, \ldots, p_{M-1}\} \tag{1}$$

$$H_T = \{t_0, t_1, \ldots, t_{L-1}\} \tag{2}$$

Here, $H_I \in \mathbb{R}^{B\times M\times d}$ and $H_T \in \mathbb{R}^{B\times L\times d}$, where B denotes the batch size, d is the feature dimension, M is the length of the image token sequence, and L is the length of the text token sequence. p_0 and t_0 are the global $[\mathrm{CLS}]$ representations for the image and text modalities, respectively; $p_1, \ldots, p_{M-1}$ correspond to image features at different spatial locations, and $t_1, \ldots, t_{L-1}$ are contextualized representations of text tokens.

The original CLIP emphasizes coarse-grained semantic consistency through global image–text alignment, but does not explicitly model fine-grained local correspondences. In cultural-heritage retrieval, however, visually similar artifacts often differ in subtle local cues. Therefore, EGF-CHITR introduces entity-prior-guided text reweighting and bidirectional local alignment on top of H_I and H_T to enhance fine-grained semantic discrimination.

B. Entity-Prior Gating Module

In cultural-heritage descriptions, key attributes such as shape, glaze type, decorative motif, craftsmanship trace, and material cue often provide visually discriminative evidence. The entity-prior gating module therefore uses domain-structured priors to adaptively increase the saliency of such tokens while suppressing redundant background narratives.

Given the contextual representations of an input text sequence, $H_T = \{t_0, t_1, \ldots, t_{L-1}\}$, we use the explicit entity positions provided by the dataset to construct a domain-entity indicator sequence $r = \{r_0, r_1, \ldots, r_{L-1}\}$ for each token. If the i-th token lies within a predefined domain entity span, we set $r_i = 1$; otherwise, $r_i = 0$. In particular, we set $r_0 = 0$ for the $[\mathrm{CLS}]$ token.

The key idea of this module is to dynamically predict a scalar importance weight g_i for each token. Specifically, we concatenate the token feature t_i with its entity indicator r_i to form an augmented vector $u_i = [t_i; r_i]$. We then compute the gating coefficient $g_i \in (0, 1)$ using a linear layer followed by a sigmoid function:

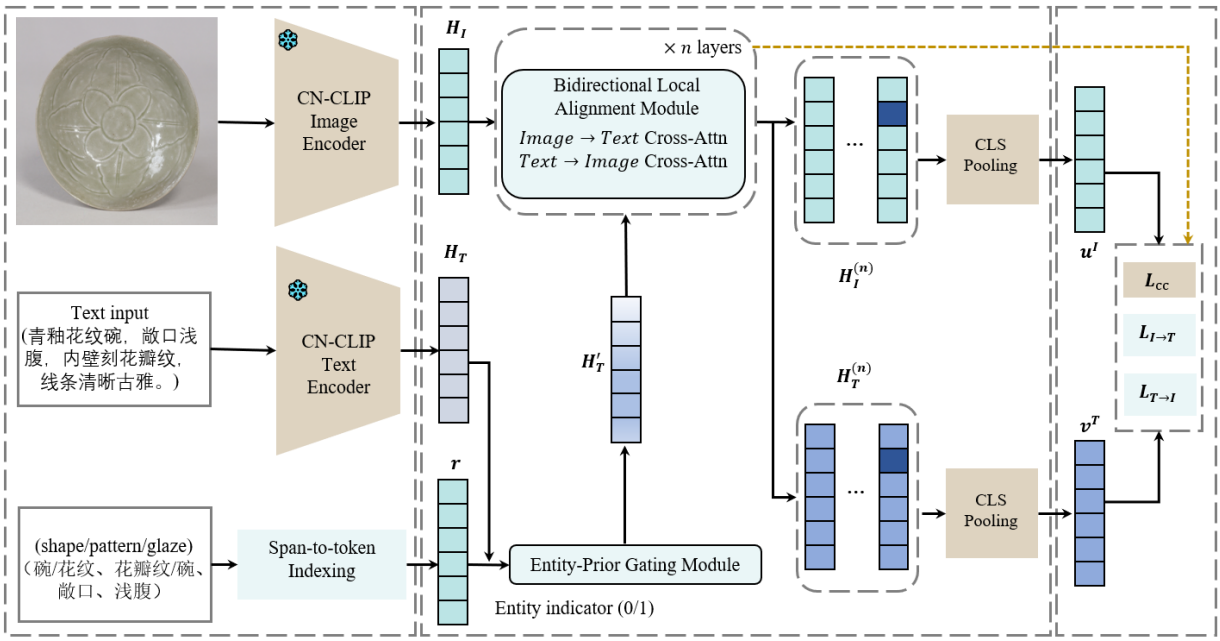

Fig. 1. Overall architecture of EGF-CHITR.

$$g_i = \sigma(W_g u_i + b) \tag{3}$$

where W_g and b are learnable parameters. Intuitively, when $r_i = 1$, i.e., the token corresponds to key entities such as "Meiziqing" glaze or "lotus-petal" patterns, the network tends to assign a larger weight to emphasize semantic cues that are strongly correlated with visual details; in contrast, auxiliary descriptions such as historical background or collection information are assigned lower weights. Finally, we adopt an entity-conditioned residual reweighting strategy to generate structure-aware text representations:

$$t'_i = (1 + r_i g_i) \cdot t_i \tag{4}$$

As shown in Eq. (4), non-entity tokens ($r_i = 0$) remain unchanged, whereas entity tokens ($r_i = 1$) are enhanced according to their discriminative importance, yielding the structure-aware sequence H'_T. This design suppresses semantic asymmetry caused by redundant narratives and provides more discriminative semantic anchors for subsequent bidirectional local alignment.

C. Bidirectional Local Alignment Module

After obtaining the image token sequence $H_I^0 = H_I$ and the entity-reweighted text token sequence $H_T^0 = H'_T$, we stack n layers of a bidirectional local alignment module on top of CN-CLIP. This module explicitly models fine-grained correspondences between image patch tokens and text tokens. Its architecture is illustrated in Fig. 2. Let the inputs of the l-th layer be H_I^{l-1} and H_T^{l-1}. Each layer consists of two sequential interaction stages and outputs two directional attention weight matrices for the consistency regularization.

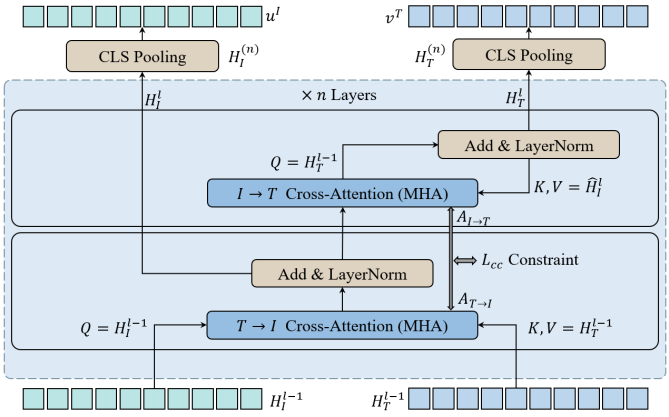

Fig. 2. Bidirectional local alignment module.

a) Text-guided image update: We compute cross-modal attention by using layer-normalized image tokens as Query and text tokens as Key and Value, producing the updated features $\hat{H}_I^l$ and the attention matrix $A_{T\to I}^l$:

$$\begin{aligned}\left(\hat{H}_I^l,\, A_{T\to I}^l\right) = \mathrm{MHA}_{T\to I}\Big(&\mathrm{LN}(H_I^{l-1}),\\ &\mathrm{LN}(H_T^{l-1}),\, \mathrm{LN}(H_T^{l-1})\Big)\end{aligned} \tag{5}$$

where $A_{T\to I}^l \in \mathbb{R}^{B\times h\times M\times L}$ denotes the attention weights when image tokens aggregate information from text tokens, h is the number of attention heads, B is the batch size, M is the image token length, and L is the text token length. We then update the intermediate image representation via a residual connection:

$$\tilde{H}_I^l = H_I^{l-1} + \hat{H}_I^l \tag{6}$$

b) Image-guided text update: We compute reverse cross-modal attention by using layer-normalized text tokens as Query and the updated intermediate image representation $\tilde{H}_I^l$ as Key and Value, producing $\hat{H}_T^l$ and the attention matrix $A_{I\to T}^l$:

$$\left(\hat{H}_T^l,\, A_{I\to T}^l\right) = \mathrm{MHA}_{I\to T}\Big(\mathrm{LN}(H_T^{l-1}),\, \mathrm{LN}(\tilde{H}_I^l),\, \mathrm{LN}(\tilde{H}_I^l)\Big) \tag{7}$$

where $A_{I\to T}^l \in \mathbb{R}^{B\times h\times L\times M}$ measures how each text token attends to image tokens. The final outputs of this layer are updated as:

$$H_T^l = H_T^{l-1} + \hat{H}_T^l \tag{8}$$

$$H_I^l = \tilde{H}_I^l \tag{9}$$

c) Attention consistency regularization: To encourage logically consistent semantic focus across the two directions, we compute an attention consistency regularization loss L_{cc} using the extracted attention matrices. For the l-th layer, we compute the mean squared error between $A_{T\to I}^l$ and the transpose of $A_{I\to T}^l$ along the last two dimensions, averaged over all samples and heads:

$$L_{\mathrm{cc}}^l = \frac{1}{B\,h\,M\,L}\sum_{b=1}^{B}\sum_{k=1}^{h}\left\| A_{T\to I}^{l,b,k} - \left(A_{I\to T}^{l,b,k}\right)^\top \right\|_F^2 \tag{10}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The overall consistency loss is the average across n layers:

$$L_{\mathrm{cc}} = \frac{1}{n}\sum_{l=1}^{n} L_{\mathrm{cc}}^l \tag{11}$$

This term is added to the total optimization objective to improve the robustness of fine-grained alignment.

After n interaction layers, we take the [CLS] token from the final-layer sequences as the global embeddings:

$$u_I = H_I^n(:,0,:) \tag{12}$$

$$v_T = H_T^n(:,0,:) \tag{13}$$

They are further L_2-normalized and used to compute the final contrastive learning objective.

D. Contrastive Learning Objective

Using the final-layer global embeddings defined above, we compute the contrastive retrieval objective as follows.

These embeddings are further L_2-normalized and used for similarity computation. For a mini-batch of size B, let the normalized image and text embeddings be $\{\mu_i^I\}_{i=1}^B$ and $\{\nu_i^T\}_{i=1}^B$, respectively. We compute the cross-modal similarity score as

$$S_{ij} = \alpha\left(\mu_i^I\right)^\top \nu_j^T, \tag{14}$$

where α is a learnable logit scale. Following CLIP, we parameterize it as

$$\alpha = \exp(\gamma) = \frac{1}{\tau}, \tag{15}$$

where γ is the optimized parameter and τ is the corresponding temperature. In implementation, we initialize $\tau = 0.07$, equivalently $\gamma = \log(1/0.07)$, and optimize γ during training. We additionally clamp the effective logit scale for numerical stability.

We adopt the symmetric InfoNCE loss:

$$L_{I\to T} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ij})}, \tag{16}$$

$$L_{T\to I} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ji})}. \tag{17}$$

The final objective is defined as

$$L = \frac{1}{2}\left(L_{I\to T} + L_{T\to I}\right) + \lambda L_{\mathrm{cc}}, \tag{18}$$

where λ controls the contribution of the attention consistency regularizer. In our experiments, λ is set to 0.1 according to validation performance.

Compared with the original CLIP objective, the proposed formulation does not change the global contrastive training principle, but injects fine-grained structure before global similarity computation through entity-prior-guided text reweighting and bidirectional patch–token interaction. This allows the final image and text embeddings to encode more discriminative local evidence while preserving global semantic consistency.

IV. Experimental Setup and Results

A. Datasets

To evaluate EGF-CHITR on fine-grained cultural-heritage image–text retrieval, we conduct experiments on two representative datasets: the in-house ceramics dataset PorTi and the public silk/mural dataset CulTi [11].

1) Ceramics Dataset PorTi: PorTi is collected from authoritative digital repositories of cultural institutions and professional academic resources, including open collections from the Palace Museum and the Longquan Celadon Museum, as well as scholarly catalogues such as *Complete Works of Unearthed Chinese Ceramics*. After image deduplication, low-quality sample filtering, and image–text normalization, the dataset contains 17,177 high-quality image–text pairs. Compared with general-purpose datasets such as MSCOCO [28], PorTi exhibits denser domain semantics and substantially stronger fine-grained discrimination requirements. Its captions mainly describe attributes such as morphological structures, glaze types, and decorative techniques, and each sample contains at least one entity cue that can be linked to a domain knowledge graph, providing stable semantic anchors for the entity-prior gating module.

Entity spans in PorTi are annotated through a human–AI collaborative pipeline. In the pre-annotation stage, an

LLM is used to extract entity spans and generate preliminary category and position labels; in the human verification stage, two experts refine the annotations according to a unified guideline, focusing on terminology normalization, boundary correction, and category rectification. On 5% independently double-annotated samples, Cohen's κ reaches 0.783, indicating strong inter-annotator agreement. We split the dataset into training/validation/test sets with a ratio of 8:1:1 at the artifact-instance level to reduce information leakage across splits.

2) Silk and Mural Dataset CulTi: To evaluate cross-category transferability, we further conduct experiments on CulTi, a public cultural-heritage dataset covering artifact categories such as silk and murals. Unlike PorTi, CulTi does not provide expert-annotated entity spans. We therefore construct weaker structured priors using a lightweight LLM-assisted extraction pipeline: for each description, the LLM identifies visually relevant technical expressions related to artifact category, material, craftsmanship, pattern, and decorative cues; the extracted candidates are then normalized and aligned to token spans to form the binary entity-indicator sequence used by the entity-prior gating module. Although these priors are weaker and noisier than those in PorTi, they still provide useful structural cues for evaluating the robustness and transferability of EGF-CHITR across heterogeneous cultural-heritage domains.

B. Evaluation Metrics

We evaluate retrieval performance using Top-K Recall (R@K) and mean Recall (mR). R@K measures the percentage of queries whose ground-truth match appears within the top-K retrieved results. Following common practice, we report R@1, R@5, and R@10 for $K \in \{1, 5, 10\}$. The mean Recall (mR) is computed as the arithmetic mean of these three recall scores. Higher R@K and mR values indicate better retrieval performance.

C. Experimental Settings and Implementation Details

All experiments are conducted on a single server with an Intel i9-13900K CPU, an NVIDIA RTX 4090 GPU (24 GB), and 64 GB RAM. We implement EGF-CHITR in Python 3.9 with PyTorch 2.0.1, using CN-CLIP as the backbone. Unless otherwise specified, ViT-B/16 and RoBERTa are adopted as the visual and text encoders, respectively, both initialized from public pre-trained weights.

During training, we freeze both backbone encoders and optimize only the lightweight task-specific modules. This frozen-backbone setting is adopted as a parameter-efficient training choice under a unified bi-encoder retrieval protocol, rather than as a strict architectural requirement of EGF-CHITR. Full fine-tuning or LoRA/Adapter-based variants are left for future work.

We train the model for 20 epochs with a batch size of 128 using Adam with an initial learning rate of 1×10^{-4}. A 5% linear warmup schedule is followed by cosine annealing, and dropout is set to 0.1. According to the ablation study, the default number of stacked interaction layers in the bidirectional local alignment module is set to $n = 3$. The consistency weight λ in the final loss is set to 0.1.

All methods are evaluated under a bi-encoder retrieval setting based on embedding matching. Specifically, the final normalized image and text embeddings are used to compute retrieval similarity, without using cross-encoder reranking or generative reranking pipelines. This keeps the compared methods as computationally comparable as possible under a unified retrieval protocol.

Following the standard CLIP parameterization, the learnable logit scale is initialized via $\tau = 0.07$, equivalently $\alpha = \exp(\gamma) = 1/\tau$, and optimized during training.

D. Experimental Results

We conduct experiments on PorTi and CulTi and compare EGF-CHITR with a representative set of recent baselines covering three categories: (1) representative bi-encoder retrieval baselines, including CN-CLIP [16]; (2) frozen-backbone adaptation and fine-grained interaction methods, including CoCoOp [18] and SCAN [12]; and (3) inference-time local matching or fine-grained enhanced retrieval methods, including Pic2Word [29], LACLIP [11], and FineCLIP [30]. We mainly select methods that can be fairly compared under the same or closely related retrieval protocol. In particular, our evaluation focuses on bi-encoder retrieval, so stronger cross-encoder or generative reranking pipelines are not included here, as they belong to a different retrieval paradigm and involve substantially different computational budgets. Nevertheless, extending the comparison to more recent large-scale retrieval and reranking methods remains important future work.

Table I summarizes the image–text retrieval results on PorTi and CulTi. Overall, EGF-CHITR achieves the best performance in both Text2Img and Img2Text directions, showing that entity-prior gating and bidirectional local alignment consistently improve fine-grained cultural-heritage retrieval. On PorTi, EGF-CHITR improves over the CN-CLIP baseline by 7.1, 8.7, and 8.7 percentage points in Text2Img R@1, R@5, and R@10, respectively, and by 2.6, 3.8, and 4.4 points in Img2Text. Compared with CoCoOp and Pic2Word, it remains superior across all R@K metrics, indicating that prompt tuning alone or one-way pseudo-token injection is insufficient to capture locally discriminative cultural-heritage cues.

EGF-CHITR also yields consistent gains on CulTi, suggesting that the proposed framework remains effective when motifs and textures are dense but textual descriptions are more globally summarized. Although PorTi provides stronger and cleaner entity priors through expert verification, the consistent gains on CulTi indicate that the improvement is not solely due to high-quality manual supervision, but also to the explicit coupling of token-level structured priors with bidirectional local alignment. On CulTi, EGF-CHITR achieves an Img2Text R@1 of 24.7, outperforming LACLIP (23.6) and FineCLIP (20.8). This observation suggests that although inference-time weighted matching or generic fine-grained enhanced pretraining can bring improvements, explicitly incorporating

TABLE I
COMPARATIVE EXPERIMENTS ON THE PORTI AND CULTI DATASETS

Method	PorTi						CulTi					
	Text2Img			Img2Text			Text2Img			Img2Text		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CN-CLIP	21.4	49.1	58.4	20.6	47.7	57.0	19.7	43.9	54.8	17.4	40.9	53.0
CoCoOp	21.9	50.9	60.3	21.8	48.5	58.7	20.7	45.2	56.1	18.1	42.0	54.4
SCAN	23.5	52.0	62.1	22.4	49.6	59.8	23.2	51.5	63.4	20.5	47.3	58.9
Pic2Word	24.0	54.4	63.4	22.6	50.0	60.2	25.2	54.8	65.2	22.1	48.9	60.3
LACLIP	27.0	57.5	66.8	23.1	51.0	61.2	28.1	56.6	66.2	23.6	49.9	62.9
FineCLIP	23.8	52.8	63.0	22.9	50.5	60.7	23.2	52.1	63.7	20.8	47.1	58.6
EGF-CHITR (Ours)	**28.5**	**57.8**	**67.1**	**23.2**	**51.5**	**61.4**	**29.1**	**57.1**	**67.3**	**24.7**	**50.1**	**63.2**

structured priors and modeling bidirectional local interactions provides more direct and stable benefits for fine-grained cultural-heritage retrieval.

E. Ablation Studies

1) Analysis of the entity-prior gating module: To verify the contribution of the entity-prior gating module to fine-grained alignment, we design three variants on the PorTi dataset for comparative analysis:

(1) No-ERW. Remove entity-guided reweighting and feed raw text features H_T to the bidirectional local alignment module.

(2) Span-ERW. Reweight using span positions only (no entity-category priors).

(3) Full. Use expert-annotated entities as entity-span priors for reweighting.

TABLE II
ABLATION RESULTS OF THE ENTITY-PRIOR GATING MODULE

Variant	Text2Img		Img2Text	
	R@1	mR	R@1	mR
No-ERW	27.2	49.9	23.0	44.6
Span-ERW	27.9	50.5	**23.4**	45.1
EGF-CHITR	**28.5**	**51.1**	23.2	**45.4**

Results in Table II show that the entity-prior gating module improves Text2Img R@1 from 27.2 to 28.5, indicating that explicitly highlighting key attribute terms can alleviate semantic asymmetry caused by redundant background narratives and guide cross-modal interaction toward visually relevant tokens. "Span-ERW" already improves over "No-ERW", suggesting that span positions alone provide useful structural cues for fine-grained alignment. Adding expert-annotated entity priors further improves the overall Text2Img performance and mean recall, although Img2Text R@1 shows a slight fluctuation. This suggests that stronger entity priors mainly provide more stable overall alignment benefits, rather than uniform gains on every metric.

2) Analysis of the bidirectional local alignment module: To investigate the key design choices of the bidirectional local alignment module, we take the full model as the reference and construct a set of variants on the PorTi dataset:

(1) No-CMI. Remove the bidirectional local alignment module.

(2) $\mathrm{I}\rightarrow\mathrm{T}$ **Only.** Keep only a single interaction direction, i.e., the image-guided text update.

(3) Bi-Dir. Use bidirectional interaction while varying the number of stacked layers $n \in \{1, 2, 4\}$; our model corresponds to $n = 3$.

(4) CC-Reg. Remove the attention consistency regularization loss L_{cc}.

TABLE III
ABLATION RESULTS OF THE BIDIRECTIONAL LOCAL ALIGNMENT MODULE

Variant	n	Direction	L_{cc}	mR	
				Text2Img	Img2Text
No-CMI	–	–	–	46.2	41.9
$\mathrm{I}\rightarrow\mathrm{T}$ Only	3	Uni	No	47.0	43.2
Bi-Dir(1)	1	Bi	Yes	49.4	44.2
Bi-Dir(2)	2	Bi	Yes	50.3	45.0
EGF-CHITR	3	Bi	Yes	**51.1**	**45.4**
Bi-Dir(4)	4	Bi	Yes	50.0	44.6
CC-Reg	3	Bi	No	50.6	45.0

Results in Table III show that any form of cross-modal interaction outperforms the no-interaction setting, and bidirectional interaction further improves over the unidirectional variant, indicating that mutual contextual perception helps alleviate semantic mismatch in ceramics retrieval. Removing L_{cc} leads to performance drops in both retrieval directions, suggesting that the consistency regularizer stabilizes bidirectional local correspondence modeling rather than merely serving as an auxiliary penalty. Performance improves from $n = 1$ to $n = 3$ and then saturates at $n = 4$, so we set $n = 3$ by default considering both effectiveness and computational cost.

V. CONCLUSION AND FUTURE WORK

This paper investigates fine-grained image–text retrieval for cultural heritage. We construct PorTi, a ceramics dataset with expert-verified entity spans, and propose EGF-CHITR, a lightweight retrieval framework built on a frozen CN-CLIP backbone. By combining an entity-prior gating module with a bidirectional local alignment module, EGF-CHITR improves the modeling of visually discriminative attributes and

proposal-free patch–token interaction. Extensive experiments on PorTi and CulTi demonstrate that the proposed framework consistently improves retrieval accuracy and robustness in both retrieval directions.

Nevertheless, the current study still has several limitations. The method depends on the quality of structured priors and may become less effective when entity spans are noisy, incomplete, or weakly grounded in visual evidence. In addition, the bidirectional cross-attention introduces extra computational overhead compared with purely global matching. Moreover, the present evaluation mainly focuses on cultural-heritage retrieval, and its effectiveness on broader general-domain benchmarks remains to be further verified. In future work, we will explore weaker-supervision settings, more efficient sparse interaction mechanisms, broader cross-domain evaluations, and extensions toward multimodal cultural-heritage knowledge organization and knowledge-graph construction.

ACKNOWLEDGMENT

This work was supported by the National Social Science Fund of China under Grant 22BMZ038.

REFERENCES

[1] M. Fiorucci, M. Khoroshiltseva, M. Pontil, A. Traviglia, A. Del Bue, and S. James, "Machine learning for cultural heritage: A survey," *Pattern Recognition Letters*, vol. 133, pp. 102–108, 2020.

[2] F. Colace, R. Gaeta, A. Lorusso, M. Pellegrino, and D. Santaniello, "New AI challenges for cultural heritage protection: A general overview," *Journal of Cultural Heritage*, vol. 75, pp. 168–193, 2025.

[3] J. Li, X. Zheng, I. Watanabe, and Y. Ochiai, "A systematic review of digital transformation technologies in museum exhibition," *Computers in Human Behavior*, vol. 161, p. 108407, 2024.

[4] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8927–8948, 2022.

[5] S. Prasomphan, "Toward fine-grained image retrieval with adaptive deep learning for cultural heritage image," *Computer Systems Science & Engineering*, vol. 44, no. 2, 2023.

[6] C. Santini, E. Posthumus, T. Tietz, M. A. Tan, O. Bruns, and H. Sack, "Multimodal search on Iconclass using vision-language pre-trained models," in *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2023, pp. 285–287.

[7] R. Huang, Y. Long, J. Han, H. Xu, X. Liang, C. Xu, and X. Liang, "NLIP: Noise-robust language-image pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023, pp. 926–934.

[8] Z. Pan, F. Wu, and B. Zhang, "Fine-grained image-text matching by cross-modal hard aligning network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 275–19 284.

[9] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improved visual-semantic embeddings," *arXiv preprint arXiv:1707.05612*, 2017.

[10] R. Xiao, S. Kim, M.-I. Georgescu, Z. Akata, and S. Alaniz, "FLAIR: VLM with fine-grained language-informed image representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 24 884–24 894.

[11] J. Yuan, J. Zhang, F. Wu, H. Lu, D. Lu, and Q. Wang, "Towards cross-modal retrieval in chinese cultural heritage documents: Dataset and solution," in *International Conference on Document Analysis and Recognition*, 2025, pp. 570–586.

[12] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[14] N. Garcia and G. Vogiatzis, "How to read paintings: Semantic art understanding with multi-modal retrieval," in *Proceedings of the European Conference on Computer Vision Workshops (ECCV Workshops)*, 2018.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[16] A. Yang, J. Pan, J. Lin, R. Men, Y. Zhang, J. Zhou, and C. Zhou, "Chinese CLIP: Contrastive vision-language pretraining in chinese," *arXiv preprint arXiv:2211.01335*, 2022.

[17] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 816–16 825.

[19] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *International Conference on Machine Learning (ICML)*, 2019, pp. 2790–2799.

[20] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[22] J. Li, R. R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in Neural Information Processing Systems*, 2021, pp. 9694–9705.

[23] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 3208–3216.

[24] M. Stefanini, M. Cornia, L. Baraldi, M. Corsini, and R. Cucchiara, "Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain," in *International Conference on Image Analysis and Processing*, 2019, pp. 729–740.

[25] W. Lin, J. Chen, J. Mei, A. Coca, and B. Byrne, "Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering," in *Advances in Neural Information Processing Systems*, 2023, pp. 22 820–22 840.

[26] X. Jiang, H. Tang, J. Gao, X. Du, S. He, and Z. Li, "Delving into multimodal prompting for fine-grained visual classification," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024, pp. 2570–2578.

[27] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "RegionCLIP: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 793–16 803.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[29] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Pic2Word: Mapping pictures to words for zero-shot composed image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 305–19 314.

[30] D. Jing, X. He, Y. Luo, N. Fei, G. Yang, W. Wei, H. Zhao, and Z. Lu, "FineCLIP: Self-distilled region-based CLIP for better fine-grained understanding," in *Advances in Neural Information Processing Systems*, 2024, pp. 27 896–27 918.