LLM-Driven Completeness and Consistency Evaluation for Cultural Heritage Data Augmentation in Cross-Modal Retrieval

Jian Zhang[1†] Junyi Guo[1†], Junyi Yuan[1], Huanda Lu[2], Yanlin Zhou[3], Fangyu Wu[1*], Qiufeng Wang[1], Dongming Lu[4]

[1]Xi'an Jiaotong-Liverpool University, [2]NingboTech University, [3]Dunhuang Academy, [4]Zhejiang University

Correspondence: fangyu.wu02@xjtlu.edu.cn

Abstract

Cross-modal retrieval is essential for interpreting cultural heritage data, but its effectiveness is often limited by incomplete or inconsistent textual descriptions, caused by historical data loss and the high cost of expert annotation. While large language models (LLMs) offer a promising solution by enriching textual descriptions, their outputs frequently suffer from hallucinations or miss visually grounded details. To address these challenges, we propose C^3, a data augmentation framework that enhances cross-modal retrieval performance by improving the completeness and consistency of LLM-generated descriptions. C^3 introduces a completeness evaluation module to assess semantic coverage using both visual cues and language-model outputs. Furthermore, to mitigate factual inconsistencies, we formulate a Markov Decision Process to supervise Chain-of-Thought reasoning, guiding consistency evaluation through adaptive query control. Experiments on the cultural heritage datasets CulTi and TimeTravel, as well as on general benchmarks MSCOCO and Flickr30K, demonstrate that C^3 achieves state-of-the-art performance in both fine-tuned and zero-shot settings. The code of this paper is available at https://github.com/JianZhang24/C-3.

1 Introduction

Cultural heritage reflects the historical, artistic, and social dimensions of human civilization across regions and periods (Nilson and Thorell, 2018). As traditional cultural heritage often comprises abstract images or patterns, textual descriptions are crucial for connecting visual data to meaningful cultural interpretations. In this paper, we focus on cross-modal retrieval tasks that enable effective matching and interpretation of cultural heritage data. This capability is fundamental for real-world applications such as digital preservation and interactive museum systems.

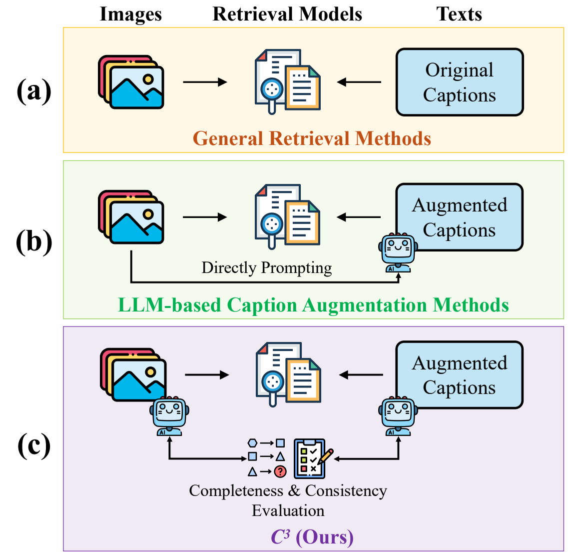


Figure 1: Illustration of our C^3 with the general retrieval methods and the LLM-based caption augmentation methods.

Cross-modal retrieval, which aims to retrieve relevant samples in one modality given a query from another, has been extensively studied on general datasets such as MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014). Recent advances have achieved significant improvements by enhancing representation learning and designing fine-grained matching strategies (Radford et al., 2021; Huang et al., 2024; Yang et al., 2024b). Nevertheless, cross-modal retrieval performance can degrade significantly when textual descriptions lack accuracy, completeness, or sufficient detail (Ma et al., 2025; Wang et al., 2025a; Sogi et al., 2024). This issue is even more pronounced in the cultural heritage domain, where historical data loss leads to incomplete records, and producing reliable annotations often requires interdisciplinary expertise. The resulting sparsity and inconsistency in textual

[†]Equal contribution.
[*]Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19419–19429
November 4-9, 2025 ©2025 Association for Computational Linguistics

descriptions substantially hinder the effectiveness of multimodal retrieval in this context.

In recent years, the rapid development of large language models (LLMs) (Guo et al., 2025; Hurst et al., 2024; Yang et al., 2024a) has led to significant advancements in enhancing textual descriptions (Wan et al., 2024; Wu et al., 2024c; Hu et al., 2024). Inspired by these advances, we explore LLM-based text augmentation methods to address the limited completeness and fidelity of textual description, thereby improving cross-modal retrieval performance. To achieve this goal, two key challenges arise: (1) how to ensure that the augmented textual annotations comprehensively capture the relevant image content; (2) how to ensure the factual accuracy of generated descriptions by addressing hallucinations introduced by LLMs.

In this paper, we propose C^3, a LLM-driven data augmentation framework designed to enhance **C**ross-modal retrieval in cultural heritage domains by improving the **C**ompleteness and **C**onsistency of textual descriptions. As shown in Fig. 1(a), general retrieval methods rely on feature alignment between the image and the original caption. LLM-augmented approaches (Fig. 1(b)) instead enhance the original caption, but often overlook whether each generated detail is grounded in visual evidence, leading to hallucinated or incomplete descriptions. To address this limitation, C^3 enhances supervision by augmenting descriptions and explicitly validating their completeness and factual correctness to reduce hallucination, thereby improving retrieval performance.

For completeness, C^3 introduces a bidirectional coverage attention evaluation approach that ensures both visual and textual information are mutually verified, establishing a cross-check mechanism to capture all relevant attributes. We further implement a coverage-based scoring method to measure how well the generated captions reflect key semantic content. This design enables our model to check that all relevant visual and textual information is captured and aligned. To improve consistency, we propose a Chain-of-Thought prompting strategy and supervise it using a Markov decision process. This supervision guides each CoT reasoning step and mitigates hallucinations by ensuring that generated textual description is grounded in visual evidence. Together, these innovations jointly enable C^3 to generate enhanced captions that are both semantically comprehensive and factually reliable, leading to improved cross-modal retrieval performance.

In summary, our contribution has four folds:

- We propose C^3, a large language model-driven data augmentation framework for cross-modal retrieval that validates generated descriptions for completeness and factual consistency. Rather than modifying retrieval model architectures, C^3 improves performance by enhancing the quality and reliability of textual supervision, enabling robust retrieval under incomplete text description.

- We introduce a novel completeness evaluation module via bidirectional coverage attention evaluation that integrates attention mechanisms with large language models, enabling comprehensive cross-verification between visual and textual modalities.

- We design a Chain-of-Thought prompting strategy supervised by a Markov decision process, effectively reducing hallucinations and improving the consistency of augmented textual description.

- Extensive experiments show that our method achieves state-of-the-art results on cultural heritage datasets CulTi and TimeTravel, as well as on general benchmarks MSCOCO and Flickr30K under both fine-tuning and zero-shot settings.

2 Related Work

2.1 Cross-Modal Retrieval

Existing cross-modal retrieval methods are typically divided into global and local alignment approaches. Global alignment methods such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) employ contrastive learning to align image-text pairs at the representation level. Chinese-CLIP (Yang et al., 2022) extends these approaches specifically to Chinese datasets, addressing linguistic and cultural specificities. In comparison, local alignment methods focus on fine-grained associations between image regions and corresponding textual elements. For example, LexVLA (Li et al., 2024b) improves interpretability and reduces false matches by introducing an overuse penalty mechanism. Similarly, GOAL (Choi et al., 2025) leverages the segmentation model to achieve more accurate local alignment. Recent studies (Chen et al., 2023; Liu et al., 2024; Pan et al., 2023; Wu et al., 2024a) have shown that incomplete or underspecified textual descriptions pose challenges for cross-modal retrieval. This issue is further exacerbated in cultural heritage domain, which typically lack detailed and accurate descriptions. In this paper, we explore the caption augmentation for improving cross-modal retrieval performance.

2.2 Large Language Models-based Data Augmentation

Recent advances in LLMs have enhanced data augmentation by enabling the generation of high-quality textual descriptions, which helps address annotation scarcity and improves generalization (Wu et al., 2024d; Fan et al., 2023). PhiloGPT (Zhang et al., 2024b) and DAR (Song et al., 2024) demonstrate the use of LLM-based data augmentation in specific domains to address annotation scarcity. However, LLMs often exhibit "hallucination", producing content that appears plausible but lacks a factual basis (Huang et al., 2025), leading to inaccurate or incomplete descriptions. To address this issue, Alizadeh et al. (2023) propose a reference-based comparison with human annotations to quantify output quality and filter unreliable samples. DoAug (Wang et al., 2025b) addresses hallucination by fine-tuning models on high-quality, validated datasets. While such methods have demonstrated effectiveness in reducing hallucination, they often incur high labor costs and risk overfitting. In this paper, we aim to improve both the completeness and consistency of LLM-augmented textual descriptions, while avoiding costly human supervision and retaining scalability across domains.

3 Motivation

Recent advances in LLM-based data augmentation have motivated us to explore their potential for improving cross-modal retrieval in the cultural heritage domain. In this section, we first investigate the necessity of evaluating the completeness of augmented captions through LLM in Section 3.1, and then discuss the role of Markov decision in enhancing the consistency of CoT-based caption augmentation in Section 3.2.

3.1 Completeness Evaluation for LLM-Augmented Descriptions

In cultural heritage retrieval, LLMs offer a promising solution to incomplete or partially informative descriptions by enriching them with additional semantic content for improved cross-modal alignment. However, the effectiveness of augmentation relies on generating descriptions that comprehensively capture all relevant visual attributes. One intuitive strategy is to use VLLMs to extract such attributes and assess whether they are sufficiently reflected in the text (Wang et al., 2023; Wu et al., 2024b). Although VLLMs are capable of generating semantically rich descriptions, their reliance on language priors often leads to incomplete coverage of visually grounded attributes. Conversely, attention-based methods offer more grounded supervision but often fail to capture rare or fine-grained visual elements, leading to incomplete coverage.

To address these limitations, we draw inspiration from coverage-based attention mechanisms developed for text summarization (See et al., 2017; Tu et al., 2016), which explicitly track how thoroughly source content is represented during generation. Building on this idea, we propose a completeness evaluation framework that combines VLLM-driven attribute extraction with bidirectional coverage attention evaluation. As discussed in Section 4.2, this design enables more accurate assessment of LLM-augmented descriptions and leads to better alignment between visual content and textual representations in cross-modal retrieval.

3.2 Consistency Enhancement for CoT-Based LLM Descriptions

Chain-of-Thought (CoT) prompting provides a structured way for LLMs to generate semantically rich and coherent captions, especially in contexts requiring contextual understanding. However, this step-by-step reasoning format can also expose a key limitation of LLMs: the frequent introduction of hallucinated or inaccurate content. Such hallucinations typically occur when generating contextually rich and culturally specific captions, leading to incorrect or nonexistent symbolic attributes being described. For example, a caption may incorrectly describe symbolic details not actually present in the image, severely impacting retrieval accuracy. Addressing these factual consistency issues is therefore critical to improving the accuracy of CoT-based caption augmentation.

To mitigate this issue, we aim to introduce a mechanism that can maintain logical consistency and factual alignment throughout the reasoning process. Existing approaches (Shum et al., 2023; Li et al., 2024a) often treat each generation step independently, which leads to semantic drift and factual

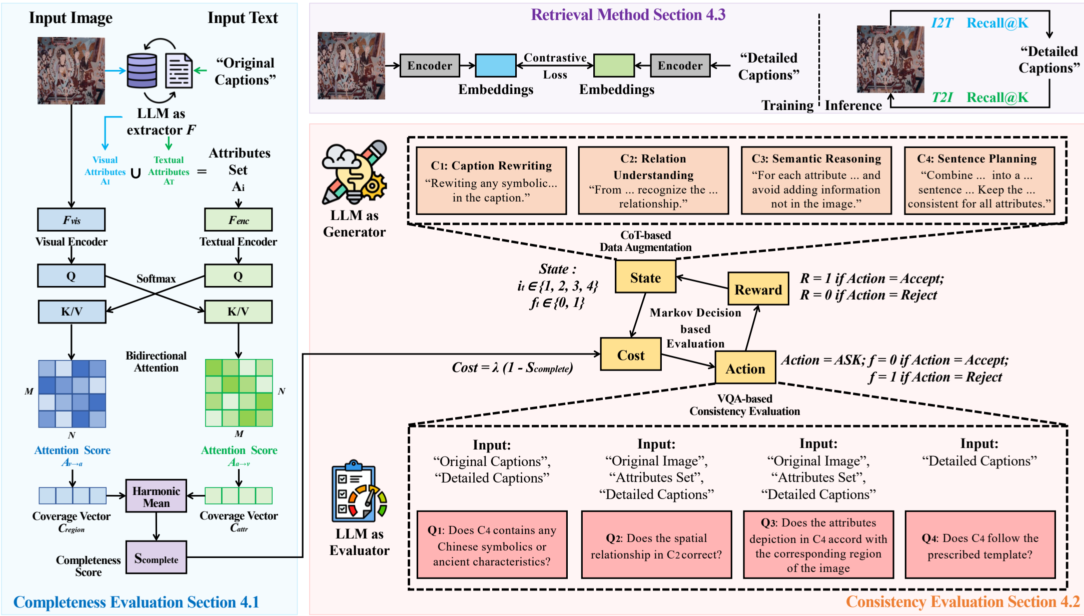


Figure 2: Overview of the proposed C^3 framework. The pipeline first verifies image-text attributes extracted by an LLM with bidirectional attention and coverage scoring, then augments captions through a CoT framework and consistency evaluation process. Detailed captions are used to fine-tune a CLIP-based retrieval model for improved image-text aligning.

errors over time. To address this, we explore the use of Markov Decision Processes (MDPs), which are effective for modeling sequential dependencies. By aligning CoT's step-by-step reasoning with the structure of MDPs, we guide the generation process more coherently and reduce hallucinations, which in turn improves factual consistency and enhances retrieval performance.

4 Methods

We propose the C^3 framework, which leverages large language models to generate complete and consistent textual descriptions, thereby enhancing cross-modal retrieval performance. The overall workflow is shown in Fig. 2.

4.1 Completeness Evaluation

Given an image–text pair (T_i, I_i), the proposed C^3 first extracts attribute-level features from both modalities. It then evaluates their cross-modal correspondence and measures coverage completeness to ensure semantic alignment.

4.1.1 Attribute Extraction

For each modality $x \in I_i, T_i$, we utilize a pre-trained multimodal large language model F to extract a set of semantic attributes:

$$A_x = F(x) = a_1^x, a_2^x, \ldots, a_{n_x}^x, \tag{1}$$

where a_j^x denotes the j-th attribute extracted from modality x, and n_x is the total number of attributes identified in x. We then merge both attribute sets into an integrated attribute pool:

$$A_i = A_{x=I_i} \cup A_{x=T_i}. \tag{2}$$

A_i provides a concise basis for subsequent completeness evaluation.

4.1.2 Bidirectional Coverage Attention Evaluation

To quantitatively evaluate the completeness of attribute coverage relative to image regions, we propose a bidirectional coverage attention evaluation. Unlike existing approaches that consider coverage from only one modality, our method simultaneously evaluates both visual-to-textual and textual-to-visual coverage. First, visual embeddings V_{I_i} for image I_i are computed using a pretrained vision encoder F_{vis} to obtain semantic embeddings:

$$V_{I_i} = F_{\text{vis}}(I_i) \in \mathbb{R}^{N \times d_v}, \tag{3}$$

where N is the number of spatial regions in the visual feature map and d_v is the visual embedding dimension.

Next, attribute embeddings $E(A_i)$ for A_i are generated via a pretrained textual encoder F_{enc}:

$$E(A_i) = F_{\text{enc}}(A_i) \in \mathbb{R}^{M \times d_t}, \tag{4}$$

where M represents the number of attributes and d_t denotes the dimension of textual embeddings. To jointly evaluate how comprehensively the visual attributes cover each spatial region and how thoroughly each attribute is grounded in visual evidence, we apply the visual regions in the image as queries while attribute embeddings as keys/values:

$$A_{\text{v}\to\text{a}} = \operatorname{softmax}\Big(\frac{V_{I_i} W^Q \big(E(A_i) W^K\big)^\top}{\sqrt{d_k}}\Big) \in \mathbb{R}^{N\times M}, \tag{5}$$

where $W^Q \in \mathbb{R}^{d_v \times d_k}$ and $W^K \in \mathbb{R}^{d_t \times d_k}$ are learnable projections.

Then, we apply attributes as queries while the visual tokens as keys/values:

$$A_{\text{a}\to\text{v}} = \operatorname{softmax}\Big(\frac{E(A_i) W^{Q'} \big(V_{I_i} W^{K'}\big)^\top}{\sqrt{d_k}}\Big) \in \mathbb{R}^{M\times N}, \tag{6}$$

where $W^{Q'}, W^{K'} \in \mathbb{R}^{d_t \times d_k}$ are another sets of learnable projections.

We calculate the maximum attention along the attended dimension yields two coverage vectors:

$$C_j^{\text{region}} = \max_{1\le m\le M} \big[A_{\text{v}\to\text{a}}\big]_{j,m}, \quad j = 1, \dots, N, \tag{7}$$

$$C_m^{\text{attr}} = \max_{1\le j\le N} \big[A_{\text{a}\to\text{v}}\big]_{m,j}, \quad m = 1, \dots, M. \tag{8}$$

Unified Completeness Score. We compute the mean coverage of image regions, $\bar{c}_{\text{region}}$, and the mean coverage of attributes, $\bar{c}_{\text{attr}}$, defined respectively as:

$$\bar{c}_{\text{region}} = \frac{1}{N}\sum_{j=1}^{N} C_j^{\text{region}}, \qquad \bar{c}_{\text{attr}} = \frac{1}{M}\sum_{m=1}^{M} C_m^{\text{attr}}. \tag{9}$$

These two averages are then combined using the harmonic mean to ensure that high completeness scores require balanced coverage in both directions. The harmonic mean specifically penalizes cases where either visual-region or attribute coverage is weak, thereby enforcing a robust evaluation.

$$S_{\text{complete}} = \frac{2\,\bar{c}_{\text{region}}\,\bar{c}_{\text{attr}}}{\bar{c}_{\text{region}} + \bar{c}_{\text{attr}}} \in [0, 1]. \tag{10}$$

The unified completeness score S_{complete} not only quantifies alignment quality but also guides decisions in subsequent consistency evaluation, dynamically modulating evaluation thoroughness based on coverage reliability.

4.2 Consistency Evaluation

4.2.1 CoT-based Data Augmentation

To enhance the consistency of cultural-heritage captions while reducing hallucinations, we propose a structured chain-of-thought (CoT) based data augmentation framework. Specifically, we employ a CoT strategy with four stages, denoted as $\{\mathrm{C}_1, \mathrm{C}_2, \mathrm{C}_3, \mathrm{C}_4\}$, to expand cultural-heritage captions while guarding against hallucination. In C^3, we use carefully designed prompts corresponding to each phase as shown in Fig. 2. By progressing from C_1 to C_4, our approach incrementally enhances the clarity, factuality, and consistency of the generated captions in a step-wise manner.

Caption rewriting C_1 Convert historical or symbolic expressions in the original textual descriptions into clear language to improve readability.

Relation Understanding C_2 Identify spatial and action-level relationships between attributes based on visual evidence, ensuring that semantic connections are grounded in the image.

Cultural semantic reasoning C_3 Construct a reasoning chain for each attribute and convert it into a factual clause, constraining the output format and length without introducing external knowledge.

Sentence planning C_4 Apply a template to ensure the accuracy of the augmented descriptions.

4.2.2 Markov Decision-based Consistency Evaluation

To ensure consistency of the captions generated at each CoT augmentation stage, we frame the entire evaluation sequence as a Markov Decision Process (MDP) that supervises which questions are asked, when, and how strictly they are evaluated. This formulation offers two key advantages. First, traditional VQA-only filters apply a fixed list of queries regardless of sample difficulty. This design often leads to hallucinations when low-confidence answers are chained together. By embedding query issuance in an MDP with query cost, our framework minimizes the hallucination risk. Second, the CoT stages follow a set order: later steps build on the earlier ones. The MDP's state explicitly tracks the current stage, ensuring that large-language models ask questions that respect this temporal dependency instead of prematurely probing later facts.

At each evaluation step t, the VLLM receives the caption generated by the CoT process. The corresponding state in the MDP process is defined as $s_t = (i_t, f_t)$, where $i_t \in \{1, 2, 3, 4\}$ indicates the current CoT stage under evaluation, and $f_t \in \{0, 1\}$ is set to 1 if any question receives a negative ("No") response, signaling regeneration. The MDP is initialized with $i_1 = 1$ and $f_1 = 0$.

The module selects one of three actions at each state. For every stage C_t, we design a series of binary verification questions for each generation stage and integrate these queries as **ASK** actions within the MDP to govern the evaluation flow. The **ASK** action poses the next verification question. If the answer is positive, the VLLM proceeds to the next stage with $f_t = 0$ remaining unchanged. If the answer is negative, it sets $f_{t+1} = 1$, triggering regeneration. The **ACCEPT** action finalizes and retains the caption; it is allowed only when all stages ($C_1 \rightarrow C_4$) have passed verification with $i_t = 4$ and $f_t = 0$. The **REJECT** action prompts the model to regenerate the caption and restart the questioning process when a negative answer is received, indicated by $f_t = 1$.

The reward function balances evaluation thoroughness against query expense. Specifically, the ASK cost is defined as $Cost(s_t) = 1 - S_{\text{complete}}$, where $S_{\text{complete}} \in [0, 1]$ measures current completeness. A binary reward is assigned only when all evaluation steps are successfully passed (**ACCEPT** after all stages); otherwise, any failed step triggers regeneration (**REJECT**) and results in zero reward. This MDP formulation explicitly manages the evaluation sequence, encouraging precise yet efficient question allocation. By dynamically adjusting evaluation thoroughness according to caption alignment quality, it effectively reduces hallucinations and ensures accurate caption generation. We denote the augmented text as T_i^{aug}, while the final retrieval target is denoted as (T_i^{aug}, I_i).

4.3 Contrastive Learning for Cross-modal Retrieval

Based on the augmented textual descriptions from previous stages, we fine-tune a Chinese-CLIP model to enhance bidirectional image-text retrieval performance. Specifically, for each pair (T_i^{aug}, I_i), the textual element T_i^{aug} is passed through pretrained text encoder to obtain its feature embedding v_i^T, whereas the visual element I_i is processed via a pretrained vision encoder to derive its feature embedding v_i^I. The resulting feature vectors v_i^T and v_i^I are projected into a shared multimodal embedding space. $S(v_i^T, v_i^I)$ denotes the cosine similarity score between the text and image embeddings. To better capture the bidirectional nature of image–text associations, a dual-objective formulation is adopted. For the image-to-text retrieval task, the training objective is formulated as:

$$L_{i2t} = -\frac{1}{Z}\sum_{i=1}^{Z} \log \frac{\exp\big(S(v_i^I,\, v_i^T)/\gamma\big)}{\sum_{k=1}^{Z} \exp\big(S(v_i^I,\, v_k^T)/\gamma\big)}, \tag{11}$$

where Z is the total number of image-text pairs, γ is a learnable temperature parameter, and $i \neq k$. Similarly, the loss for the text-to-image retrieval task, L_{t2i}, is defined in a structurally symmetric form by interchanging the image and text representations. The overall training loss is formulated as $L = (L_{i2t} + L_{t2i})/2$. By minimizing the combined contrastive loss L, the model aligns image and text embeddings in the joint multi-modal space.

5 Experiment

5.1 Experimental Settings

Implementation Details. We employed ViT-H/14 (Dosovitskiy et al., 2020) as the visual backbone and utilized RoBERTa-wwm-large (Liu et al., 2019) as the textual backbone to construct the Chinese-CLIP (Yang et al., 2022) and CLIP (Radford et al., 2021) models. The experimental trials were executed utilizing a single NVIDIA RTX A6000 GPU, endowed with 48 gigabytes of memory. The fine-tuning process was implemented with a batch size of 32, a learning rate set to 5e-5, and trained for 3 epochs. We utilize Janus-Pro 7B (Chen et al., 2025) for attribute extraction and caption generation, and Qwen-2.5VL (Bai et al., 2025) for consistency evaluation. For the inference time, C^3 takes 2.9 s per caption, while Janus-Pro with the default prompt template takes 2.2 s per caption.

Datasets and Evaluation Metrics. We perform experiments on two publicly cross-modal retrieval datasets, MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014), as well as on a domain-specific cultural dataset CulTi (Yuan et al., 2025), and a historical artifact dataset TimeTravel (Ghaboura et al., 2025). MSCOCO comprises 123,287 images, each paired with five English captions, of which 113,287 are used for training, 5,000 for validation, and 5,000 for testing. The Flickr30K dataset contains 31,783 images with

Table 1: Performance of C^3 on MSCOCO and Flickr30K datasets in zero-shot setting. Best results are in **bold**.

Methods	MSCOCO						Flickr30K					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN (Lee et al., 2018)	50.4	82.2	90.0	38.6	69.3	80.4	67.4	90.3	95.8	48.6	77.7	85.2
ViSTA (Cheng et al., 2022)	68.9	90.1	95.4	52.6	79.6	87.6	89.5	98.4	99.6	75.8	94.2	96.9
COTS (Lu et al., 2022)	69.0	90.4	94.9	52.4	79.0	86.9	90.6	98.7	99.7	76.5	93.9	96.6
LightningDoT (Sun et al., 2021)	64.6	87.6	93.5	50.3	78.7	87.5	86.5	97.5	98.9	72.6	93.1	96.1
LexLIP (Luo et al., 2023)	70.2	90.7	95.2	53.2	79.1	86.7	**91.4**	**99.2**	**99.7**	**78.4**	**94.6**	**97.1**
CLIP (Radford et al., 2021)	51.8	76.8	84.3	32.7	57.7	68.2	44.1	68.2	77.0	24.7	45.1	54.6
Long-CLIP (Zhang et al., 2024a)	62.8	85.1	91.2	46.3	70.8	79.8	53.4	77.5	85.3	41.2	64.1	72.6
C^3 (Ours)	**86.7**	**97.5**	**98.9**	**85.8**	**97.4**	**99.1**	84.5	97.0	99.0	72.9	92.0	96.1

five English descriptions each, and we allocate 29,783 images for training and 1,000 each for validation and testing. Since our method augments captions based on the original captions, we use a one-to-one image-text pairing for retrieval. CulTi includes 5,726 Chinese image text pairs spanning two categories, silk artifacts and Dunhuang murals, with 4,008 pairs for training, 573 for validation, and 1,145 for testing. TimeTravel contains 10,250 expert-verified image–text pairs from 266 cultural groups across 10 regions, designed for historical artifact analysis and cultural context understanding. We split the dataset into 7,175 samples for training, 1,025 for validation, and 2,050 for testing. Retrieval performance is evaluated using Recall@K ($R@K$), with $K \in \{1, 5, 10\}$.

5.2 Main Results

Zero-shot Retrieval. As shown in Tab. 1. C^3 achieves competitive retrieval performance in zero-shot setting for MSCOCO (Lin et al., 2014) and Flickr30K (Young et al., 2014), with accuracy exceeding 99%. As text description is less pronounced in these datasets, augmentation brings limited improvement. However, C^3 significantly improves the original CLIP baseline (Radford et al., 2021) in both public datasets. On the more challenging CulTi dataset, C^3 outperforms LACLIP (Yuan et al., 2025) by 18.2% at $R@1$ in the zero-shot setting as shown in Tab. 2. On the TimeTravel dataset, C^3 also surpasses direct prompting by an average of 2.8% at $R@1$ across both retrieval directions, confirming its effectiveness on historical artifact retrieval. By validating both visual attributes and their semantic grounding, C^3 enhances description quality and enables superior cross-modal alignment and generalization, even without domain-specific fine-tuning. The shortfall on Flickr30K is mainly due to the CLIP-based backbone. Its pre-training is less effective for fine-grained phrase grounding required by Flickr30K.

Table 2: Retrieval performance on CulTi and TimeTravel test sets. For the direct prompt, we use the default prompt template provided by Janus-Pro.

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
CulTi (zero-shot)						
LACLIP (Yuan et al., 2025)	7.6	22.9	31.1	11.0	25.1	36.2
Direct Prompt	20.6	43.6	54.2	25.4	48.3	60.1
C^3 **(Ours)**	**25.8**	**51.3**	**61.9**	**26.4**	**55.5**	**65.7**
CulTi (fine-tune)						
LACLIP (Yuan et al., 2025)	23.6	49.9	62.9	28.1	56.6	66.2
Direct Prompt	54.6	80.4	89.3	53.7	80.3	88.1
C^3 **(Ours)**	**74.1**	**93.4**	**96.4**	**77.1**	**94.6**	**97.4**
TimeTravel (zero-shot)						
Original Caption	10.6	26.0	35.3	9.8	24.9	33.8
Direct Prompt	21.0	40.8	50.5	17.5	35.3	45.0
C^3 **(Ours)**	**24.0**	**44.6**	**54.8**	**20.0**	**38.3**	**47.8**
TimeTravel (fine-tune)						
Original Caption	21.3	49.8	65.6	21.0	48.3	64.6
Direct Prompt	42.7	70.4	81.4	43.7	71.0	80.7
C^3 **(Ours)**	**46.6**	**75.2**	**86.9**	**48.4**	**76.9**	**86.3**

Cultural Heritage Domain Retrieval. As shown in Tab. 2 and in Fig. 4 (a), the directly prompting-based method outperforms LACLIP (Yuan et al., 2025) in CulTi. Our approach further improves upon directly prompting by 19.5% at $R@1$ for image-to-text retrieval in a fine-tuned setting. On the TimeTravel dataset, C^3 also surpasses direct prompting by 4.3% at $R@1$ in the fine-tuned setting, highlighting its effectiveness on historical artifact retrieval. These results demonstrate the robustness of C^3 on challenging cultural data. By enabling attribute-level bidirectional evaluation, our method achieves fine-grained alignment between visual details and textual attributes, effectively addressing the incompleteness and ambiguity common in cultural captions. Besides, Fig. 4 (b) shows that C^3 achieves similar retrieval performance across different VLLM parameter sizes, indicating that C^3 is not dependent on model scale.

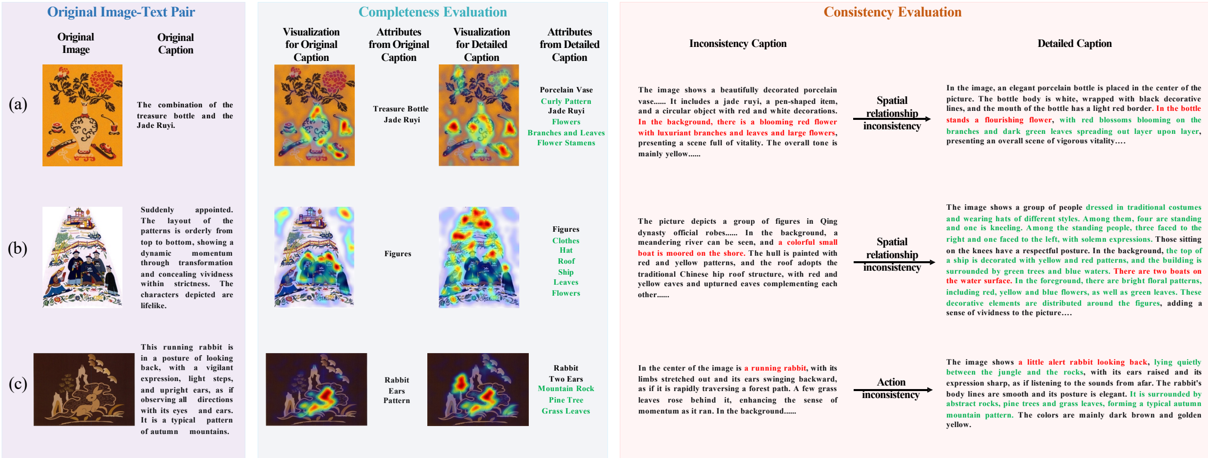
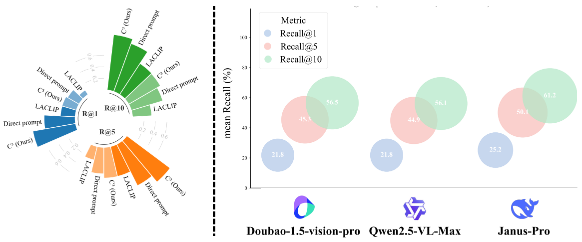


Figure 3: Case studies on completeness and consistency evaluation. The examples of captions and annotations are translated from Chinese to English for better understanding. Red denotes inaccurate or hallucinated descriptions; Green denotes missing details in the original caption.



(a) (b)

Figure 4: (a) Retrieval performance on the CulTi dataset under both zero-shot and fine-tune conditions; (b) Comparisons of zero-shot retrieval performance between different VLLMs in CulTi.

Table 3: Ablation studies on three datasets.

No.	Model Variant	Flickr30k mR	Flickr30k Rsum	MSCOCO mR	MSCOCO Rsum	CulTi mR	CulTi Rsum
(a)	C^3 **(Ours)**	95.5	572.8	93.9	563.3	88.8	532.8
(b)	w/o Completeness Evaluation	94.5	567.0	90.5	542.8	81.3	488.1
(c)	w/o CoT	94.6	567.6	88.9	533.2	74.4	446.5
(d)	w/o Consistency Evaluation	95.3	571.6	93.6	561.8	86.8	520.5

5.3 Ablation studies

5.3.1 Importance of Completeness Evaluation

We conduct an ablation study to demonstrate the effectiveness of the completeness evaluation as shown in Tab. 3 (b). To remove the completeness evaluation, we no longer extract attributes from augmentation, and each verification question is assigned an equal cost. Under these conditions, the mean Recall (mR) and sum of Recalls (Rsum) decrease by 7.5% and 44.7% on CulTi. This ablation highlights that completeness evaluation ensures that all visual regions and attributes are accounted for, thus significantly improving the representation of image content in the captions. Moreover, this step is particularly important in the cultural heritage domain, where visual attributes are diverse and frequently subtle, requiring precise completeness evaluation to support accurate retrieval.

5.3.2 Importance of Consistency Evaluation

We further evaluate the necessity of consistency evaluation. In this case, we remove the Chain-of-Thought (CoT) and Markov decision-based VQA evaluation. First, we directly generate captions using simple prompting as shown in Tab. 3 (c). In this scenario, the mR and Rsum drop by 14.4% and 86.3%, respectively. We analyze that the model cannot decompose and reason complex attribute information, resulting in a significant drop in retrieval performance. Second, we apply CoT without consistency evaluation in Tab. 3 (d). In this case, mR and Rsum drop by 2.0% and 12.3%. Additionally, our consistency evaluation yields a dialog reasoning accuracy of 93.2%, evidencing the reliability of the evaluator in guiding caption refinement. The lack of consistency evaluation increases the risk of hallucinations, especially when dealing with previously unseen cultural heritage concepts. Therefore, consistency evaluation is crucial in LLM-based data augmentation to prevent inaccuracies and ensure caption reliability in diverse cultural contexts.

5.4 Case Study

In this section, we illustrate the advantage of C^3 through three case studies in Fig. 3. These case studies focus on two typical problems that occurred during the caption augmentation processing, but could be addressed by our C^3 framework. Specifically, in Fig. 3 (a), the original and LLM-generated captions mention only the bottle and ruyi but omit the flowers at the top. Our method adds descriptions for these flowers so retrieval can account for them. By covering every visual attribute, our augmented captions form more complete queries. Therefore, the retrieval model performs stronger alignment across all described elements. In Fig. 3 (c), the rabbit's action is unclear from the image alone, but the original caption uses the word "looking back". This shows that the caption "running rabbit" is inaccurate. C^3 is able to identify these issues because the proposed consistency evaluation leverages completeness evaluation to find objective evidence from both the image and the text. Furthermore, our approach asks targeted questions at each stage of the CoT process, narrowing the scope of evaluation and enabling precise checking.

6 Conclusion

In this paper, we addressed the challenge of incomplete and unreliable textual annotations in cultural heritage retrieval through C^3, an LLM-driven data augmentation framework. By introducing a novel bidirectional validation approach and a coverage-based measure mechanism, our model ensures comprehensive textual descriptions. Furthermore, the integration of Chain-of-Thought prompting supervised by a Markov decision process effectively mitigates hallucination, significantly enhancing reliability. Empirical results validate our method's efficacy, achieving superior performance on cultural heritage retrieval tasks and demonstrating robust generalizability across benchmark datasets. Future work will explore extending this approach to additional modalities and further refining semantic alignment strategies for broader applicability.

7 Limitations

Although our proposed method C^3 has demonstrated promising performance on cultural heritage datasets, it still exhibits minor inaccuracies in generated captions. Upon manual inspection, we found occasional errors primarily arising from incomplete visual information, such as damaged murals or poor image quality, which can mislead the model's assessment of content completeness. Additionally, the consistency evaluation requires multi-turn verification for difficult cultural images, and LLMs may forget context over multiple rounds, sometimes causing incorrect captions to be mistakenly accepted as correct. Future research will address these limitations by incorporating robustness to incomplete visual contexts and by investigating dynamic context optimization strategies for the evaluation process.

Acknowledgments

This project is supported by the National Natural Science Foundation of China (Nos. 62436009, 62276258), the XJTLU Research Development Fund and Teaching Development Fund (Grant No. RDF-24-01-016, TDF23/24-R27-222), and the Suzhou Science and Technology Development Planning Programme (Grant No. ZXL2023176).

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. arXiv preprint arXiv:2307.02179.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923.

Weijing Chen, Linli Yao, and Qin Jin. 2023. Rethinking benchmarks for cross-modal image-text retrieval. In Proceedings of SIGIR, pages 1241–1251.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811.

Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, and 1 others. 2022. Vista: Vision and scene text aggregation for cross-modal retrieval. In Proceedings of CVPR, pages 5184–5193.

Hyungyu Choi, Young Kyun Jang, and Chanho Eom. 2025. Goal: Global-local object alignment learning. arXiv preprint arXiv:2503.17782.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of ICLR.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. In Proceedings of NeurIPS, pages 35544–35575.

Sara Ghaboura, Ketan More, Ritesh Thawkar, Wafa Alghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Time travel: A comprehensive benchmark to evaluate lmms on historical and cultural artifacts. arXiv preprint arXiv:2502.14865.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llm vs small model? large language model based text augmentation enhanced personality detection model. In Proceedings of the AAAI, pages 18234–18242.

Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. 2024. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In Proceedings of AAAI, pages 18298–18306.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst., 43(2).

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of ICML, pages 4904–4916.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In Proceedings of ECCV, pages 201–216.

Dailin Li, Chuhan Wang, Xin Zou, Junlong Wang, Peng Chen, Jian Wang, Liang Yang, and Hongfei Lin. 2024a. Cot-based data augmentation strategy for persuasion techniques detection. In Proceedings of ACL, pages 1315–1321.

Yifan Li, Yikai Wang, Yanwei Fu, Dongyu Ru, Zheng Zhang, and Tong He. 2024b. Unified lexical representation for interpretable visual-language alignment. In Proceedings of NeurIPS, pages 1141–1161.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In Proceedings of ECCV, pages 740–755. Springer.

Haoyu Liu, Yaoxian Song, Xuwu Wang, Xiangru Zhu, Zhixu Li, Wei Song, and Tiefeng Li. 2024. Flickr30k-cfq: A compact and fragmented query dataset for text-image retrieval. In Proceedings of DASFAA, pages 419–434.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. 2022. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In Proceedings of CVPR, pages 15692–15701.

Ziyang Luo, Pu Zhao, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Lexlip: Lexicon-bottlenecked language-image pre-training for large-scale image-text sparse retrieval. In Proceedings of ICCV, pages 11206–11217.

Zehong Ma, Hao Chen, Wei Zeng, Limin Su, and Shiliang Zhang. 2025. Multi-modal reference learning for fine-grained text-to-image retrieval. IEEE Trans. Multimedia, 27:5009–5022.

Tomas Nilson and Kristina Thorell. 2018. Cultural heritage preservation: The past, the present and the future. Halmstad University Press.

Zhengxin Pan, Fangyu Wu, and Bailing Zhang. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In Proceedings of CVPR, pages 19275–19284.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In Proceedings of ICML, pages 8748–8763.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In Proceedings of ACL, pages 1073–1083.

KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. arXiv preprint arXiv:2302.12822.

Naoya Sogi, Takashi Shibata, and Makoto Terao. 2024. Object-aware query perturbation for cross-modal image-text retrieval. In Proceedings of ECCV, pages 447–464.

Fangzhou Song, Bin Zhu, Yanbin Hao, and Shuo Wang. 2024. Enhancing recipe retrieval with foundation models: A data augmentation perspective. In Proceedings of ECCV, pages 111–127.

Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In Proceedings of ACL, pages 982–997.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In Proceedings of ACL, pages 76–85.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, and 1 others. 2024. Tnt-llm: Text mining at scale with large language models. In Proceedings of SIGKDD, pages 5836–5847.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2025a. Cross-modal retrieval: a systematic review of methods and future directions. Proceedings of the IEEE, 112:1716–1754.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and 1 others. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In Proceedings of NeurIPS, pages 61501–61513.

Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, Pengfei Wang, and Yuanchun Zhou. 2025b. Diversity-oriented data augmentation with large language models. arXiv preprint arXiv:2502.11671.

Fangyu Wu, Qiufeng Wang, Xuan Liu, Qi Chen, Yuxuan Zhao, Bailing Zhang, and Eng Gee Lim. 2024a. Discriminative feature enhancement network for few-shot classification and beyond. ESWA, 255:124811.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, and 1 others. 2024b. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. In Proceedings of NeurIPS, pages 69925–69975.

Shih-Lun Wu, Xuankai Chang, Gordon Wichern, Jee-weon Jung, François Germain, Jonathan Le Roux, and Shinji Watanabe. 2024c. Improving audio captioning models with fine-grained audio features, text embedding supervision, and llm mix-up augmentation. In Proceedings of ICASSP, pages 316–320. IEEE.

Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. 2024d. Lotlip: Improving language-image pre-training for long text understanding. In Proceedings of NeurIPS, pages 64996–65019.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

Rui Yang, Shuang Wang, Yu Gu, Jihui Wang, Yingzhi Sun, Huan Zhang, Yu Liao, and Licheng Jiao. 2024b. Continual learning for cross-modal image-text retrieval based on domain-selective attention. Pattern Recognition, 149:110273.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. of the ACL, pages 67–78.

Junyi Yuan, Jian Zhang, Dongming Lu, Huanda Lu, Fangyu Wu, and Qiufeng Wang. 2025. Towards cross-modal retrieval in chinese cultural heritage documents: Dataset and solution. arXiv preprint arXiv::2505.10921.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-clip: Unlocking the long-text capability of clip. In Proceedings of ECCV, pages 310–325.

Yuqing Zhang, Baoyi He, Yihan Chen, Hangqi Li, Han Yue, Shengyu Zhang, Huaiyong Dou, Junchi Yan, Zemin Liu, Yongquan Zhang, and Fei Wu. 2024b. PhiloGPT: A philology-oriented large language model for Ancient Chinese manuscripts with dunhuang as case study. In Proceedings of EMNLP, pages 2784–2801.