Novelty Assessment Report

Paper: RAVENEA: A Benchmark for Multimodal Retrieval-Augmented Visual Culture Understanding
PDF URL: https://openreview.net/pdf?id=4zAbkxQ23i
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

As vision-language models (VLMs) become increasingly integrated into daily life, the need for accurate visual culture understanding is becoming critical. Yet, these models frequently fall short in interpreting cultural nuances effectively. Prior work has demonstrated the effectiveness of retrieval-augmented generation (RAG) in enhancing cultural understanding in text-only settings, while its application in multimodal scenarios remains underexplored. To bridge this gap, we introduce RAVENEA (Retrieval-Augmented Visual culturE uNdErstAnding), a new benchmark designed to advance visual culture understanding through retrieval, focusing on two tasks: culture-focused visual question answering (cVQA) and culture-informed image captioning (cIC). RAVENEA extends existing datasets by integrating over 10,000 unique Wikipedia documents curated and ranked by human annotators. Through the extensive evaluation on seven multimodal retrievers and fifteen VLMs, RAVENEA reveals some undiscovered findings: (i) In general, cultural grounding annotations can enhance multimodal retrieval and corresponding downstream tasks. (ii) Lightweight VLMs, when augmented with culture-aware retrieval, outperform their non-augmented counterparts (by at least 3.2% on cVQA and 6.2% on cIC). (iii) Performance varies widely across countries, with culture-aware retrieval augmented VLMs showing more stable results on Korean and Chinese contexts than in the other countries. These findings highlight the critical limitations of current multimodal retrievers and VLMs, and underscore the need to enhance RAG visual culture understanding. Our RAVENEA can serve as a foundational tool for advancing the study of RAG visual culture understanding of multimodal AI.

Core Task Landscape

This paper addresses: **Retrieval-Augmented Visual Culture Understanding**

A total of **40 papers** were analyzed and organized into a taxonomy with **11 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Retrieval-Augmented Generation Frameworks for Vision**
- **Cultural Heritage and Artistic Understanding**
- **Cross-Cultural and Multilingual Visual Understanding**
- **Multimodal Retrieval and Benchmarking**
- **Specialized Visual Generation and Retrieval Applications**

Complete Taxonomy Tree

- Retrieval-Augmented Visual Culture Understanding Survey Taxonomy
- Retrieval-Augmented Generation Frameworks for Vision
  - General Visual RAG Architectures (3 papers)
  - [1] Retrieval augmented generation and understanding in vision: A survey and new outlook (Zheng Xu, 2025) View paper
  - [9] Video-RAG: Visually-aligned Retrieval-Augmented Long Video Comprehension (Luo Yong-dong, 2024) View paper
  - [10] RegionRAG: Region-level Retrieval-Augmented Generation for Visual Document Understanding (Yinglu Li, 2025) View paper
  - Image Captioning with Retrieval Augmentation (3 papers)
  - [3] DIR: Retrieval-Augmented Image Captioning with Comprehensive Understanding (Wu Hao, 2024) View paper
  - [11] Evcap: Retrieval-Augmented Image Captioning with External Visual-Name Memory for Open-World Comprehension (Jiaxuan Li, 2024) View paper
  - [19] Diffusion Based Augmentation for Captioning and Retrieval in Cultural Heritage (Dario Cioni, 2023) View paper
- Cultural Heritage and Artistic Understanding
  - Fine Art and Artwork Analysis (2 papers)
  - [2] ArtRAG: Retrieval-Augmented Generation with Structured Context for Visual Art Understanding (Wang, 2025) View paper
  - [25] Mosaic: Finding artistic connections across culture with conditional image retrieval (Mark Hamilton, 2021) View paper
  - Cultural Heritage Information Systems (20 papers)
  - [4] FolkRAG: a retrieval-augmented generation system for cultural heritage materials (Paul Kelly, 2025) View paper
  - [7] Multimedia information retrieval using content-based image retrieval and context link for Chinese cultural artifacts (Chung Ming Lo, 2024) View paper
  - [8] Probing historical image contexts: Enhancing visual archive retrieval through computer vision (Lin Du, 2024) View paper
  - [12] Towards Cross-Modal Retrieval in Chinese Cultural Heritage Documents: Dataset and Solution (Yuan Jun-yi, 2025) View paper
  - [14] Multi-Modal Semantic Parsing for the Interpretation of Tombstone Inscriptions (Xiao Zhang, 2025) View paper
  - [17] Immersive Virtual Museums with Spatially-Aware Retrieval-Augmented Generation (Elisa Ayumi Masasi de Oliveira, 2025) View paper
  - [18] Computer vision and AI TOOLS for enhancing user experience in the cultural heritage domain (Paolo Mazzanti, 2024) View paper

- [20] How Can Generative Artificial Intelligence Techniques Facilitate Intelligent Research into Ancient Books? (Jiangfeng Liu, 2024) View paper
- [21] Cultural heritage information retrieval: past, present, and future trends (Babak Ranjgar, 2024) View paper
- [24] When MLLMs Meet ICH: A visual retrieval-augmented generation-based method for intangible cultural heritage image recognition-take Shadow Puppetry as a case (T Fan, 2026) View paper
- [26] Vision-Augmented RAG System for Interactive Local Heritage Exploration (Ojha, 2025) View paper
- [30] Content-based image retrieval of cultural heritage symbols by interaction of visual perspectives (Kwan Ph, 2011) View paper
- [31] A review of image retrieval methods for digital cultural heritage resources (Chih Fong Tsai, 2007) View paper
- [32] Visual information retrieval from historical document images (Sara Zhalehpour, 2019) View paper
- [34] A hybrid ontology and visual-based retrieval model for cultural heritage multimedia collections (Stefanos Vrochidis, 2008) View paper
- [35] Managing digital cultural objects: analysis, discovery and retrieval (John Rodzvilla, 2016) View paper
- [36] Exploring the Integration of National Cultural Resources through Content Retrieval and Visual Semantic Segmentation Display (Huayan Zhang, 2023) View paper
- [37] Content-Based Indexing and Retrieval of Cultural Heritage Data: An Integrated Approach to Documentation with Application to the EROS (Paquet, 2006) View paper
- [38] Leading to Cultural Resources Through Visuals: An Image Retrieval System and On-site Information Approach (Kiyofumi Motoyama, 2010) View paper
- [39] Cultural Heritage Assistant: A Lightweight Retrieval Augmented Generation Method Enhanced Vision-Language Model for Cultural Heritage (Wang Shiyu, n.d.) View paper

- Cross-Cultural and Multilingual Visual Understanding
  - Cultural Alignment and Bias Evaluation (2 papers)
  - [15] CAIRe: Cultural Attribution of Images by Retrieval-Augmented Evaluation (Khanuja, 2025) View paper
  - [29] Exposing Blindspots: Cultural Bias Evaluation in Generative Image Models (Hong Minki, 2025) View paper
  - Cross-Lingual Visual Retrieval and Annotation (1 papers)
  - [13] AraTraditions10k bridging cultures with a comprehensive dataset for enhanced cross lingual image annotation retrieval and tagging (Emran Al-Buraihy, 2025) View paper
  - Culture-Specific Language Model Enhancement (3 papers)
  - [5] Valuesrag: Enhancing cultural alignment through retrieval-augmented contextual learning (Wonduk Seo, 2025) View paper
  - [6] Leveraging Retrieval-Augmented Generation for Culturally Inclusive Hakka Chatbots: Design Insights and User Perceptions (Chen-Chi Chang, 2024) View paper
  - [23] Evaluating cultural knowledge processing in large language models: a cognitive benchmarking framework integrating retrieval-augmented generation (Lee, 2025) View paper
- Multimodal Retrieval and Benchmarking
  - Visual Culture Understanding Benchmarks ★ (2 papers)
  - [0] RAVENEA: A Benchmark for Multimodal Retrieval-Augmented Visual Culture Understanding (Anon et al., 2026) View paper
  - [16] Event-Enriched Image Analysis Grand Challenge At ACM Multimedia 2025 (Tran Thien Phuc, 2025) View paper
- Specialized Visual Generation and Retrieval Applications
  - Sign Language and Accessibility (1 papers)
  - [27] Signs as Tokens: A Retrieval-Enhanced Multilingual Sign Language Generator (Zuo, 2024) View paper
  - Fashion and Social Media Content Generation (1 papers)
  - [28] FITMag: A Framework for Generating Fashion Journalism Using Multimodal LLMs, Social Media Influence, and Graph RAG (Jinda Han, 2025) View paper
  - Semantic Security and Ontology-Based Retrieval (3 papers)
  - [22] Ontology-based Secure Retrieval of Semantically Significant Visual Contents (Muhammad Khan, 2022) View paper
  - [33] LUMOS-DM: Landscape-Based Multimodal Scene Retrieval Enhanced by Diffusion Model (Viet-Tham Huynh, 2024) View paper
  - [40] A Network Perspective on Gradient Flow Equations for Deep Linear Neural Networks (J Wendin, n.d.) View paper

Narrative

Core task: retrieval-augmented visual culture understanding. This emerging field combines multimodal retrieval with cultural knowledge to interpret images, artworks, and heritage artifacts in context. The taxonomy reveals five main branches: Retrieval-Augmented Generation Frameworks for Vision develops general-purpose architectures that integrate external knowledge into vision-language models, exemplified by surveys like RAG Vision Survey[1] and systems such as Video-RAG[9] and RegionRAG[10]. Cultural Heritage and Artistic Understanding focuses on domain-specific applications for museums, historical artifacts, and traditional art forms, with works like ArtRAG[2] and Chinese Heritage Retrieval[12] addressing specialized cultural content. Cross-Cultural and Multilingual Visual Understanding tackles diversity and representation challenges through datasets like AraTraditions10k[13] and systems such as Hakka Chatbots[6]. Multimodal Retrieval and Benchmarking establishes evaluation frameworks and datasets to measure cultural comprehension capabilities, while Specialized Visual Generation and Retrieval Applications explores targeted use cases from tombstone parsing to sign language recognition.

Recent activity highlights tensions between general-purpose retrieval frameworks and culturally-grounded approaches. While broad systems like DIR Captioning[3] and Evcap[11] aim for scalable image understanding, works such as FolkRAG[4] and ValuesRAG[5] emphasize the importance of culturally-specific knowledge bases that capture nuanced traditions and values. RAVENEA[0] sits within the benchmarking cluster alongside Event Image Challenge[16], contributing evaluation resources for assessing how well models handle culturally-rich visual content. Compared to neighboring benchmarks, RAVENEA appears to emphasize retrieval-augmented approaches where external cultural knowledge enhances interpretation, contrasting with purely end-to-end vision-language evaluation. Key open questions include how to balance retrieval efficiency with cultural depth, whether general frameworks can adequately capture regional specificity, and how to construct representative benchmarks that avoid perpetuating cultural biases while enabling meaningful progress measurement.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Event-Enriched Image Analysis Grand Challenge At ACM Multimedia 2025

Authors: Tran Thien Phuc, Nguyen Minh Quang, Tran, Minh-Triet, Nguyen, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The Event-Enriched Image Analysis (EVENTA) Grand Challenge, hosted at ACM Multimedia 2025, introduces the first large-scale benchmark for event-level multimodal understanding. Traditional captioning and retrieval tasks largely focus on surface-level recognition of people, objects, and scenes, often overlooking the contextual and semantic dimensions that define real-world events. EVENTA addresses this gap by integrating contextual, temporal, and semantic information to capture the who, when, wher...

Relationship Analysis

Both papers belong to the Visual Culture Understanding Benchmarks category, focusing on datasets and evaluation protocols for assessing cultural visual understanding capabilities. While RAVENEA evaluates retrieval-augmented visual culture understanding across eight countries with human-ranked Wikipedia documents for cVQA and cIC tasks, EVENTA focuses on event-level multimodal understanding through news imagery, emphasizing contextual storytelling with event-enriched image retrieval and captioning tasks. The key difference is that RAVENEA targets cultural nuances and traditions across diverse countries, whereas EVENTA centers on temporal and narrative context within news events and journalism applications.

Contributions Analysis

Overall novelty summary. The paper introduces RAVENEA, a benchmark for retrieval-augmented visual culture understanding, comprising culture-focused visual question answering and culture-informed image captioning tasks with over 10,000 curated Wikipedia documents. Within the taxonomy, it resides in the 'Visual Culture Understanding Benchmarks' leaf under 'Multimodal Retrieval and Benchmarking'. This leaf contains only two papers total, including RAVENEA and one sibling (Event Image Challenge), indicating a relatively sparse research direction. The broader taxonomy encompasses 40 papers across multiple branches, suggesting that while cultural visual understanding is active, dedicated benchmarking efforts remain limited.

The taxonomy reveals that RAVENEA's parent branch (Multimodal Retrieval and Benchmarking) is distinct from neighboring areas like 'Cultural Heritage Information Systems' (20 papers) and 'Retrieval-Augmented Generation Frameworks for Vision' (6 papers). While heritage systems focus on organizing artifacts and RAG frameworks develop general architectures, RAVENEA bridges these by providing evaluation infrastructure specifically for retrieval-augmented cultural interpretation. The scope note clarifies that general visual benchmarks without cultural focus belong elsewhere, positioning RAVENEA at the intersection of cultural grounding and multimodal retrieval assessment. This placement suggests the work addresses an underserved niche between application-focused heritage systems and evaluation-focused general benchmarks.

Among 30 candidates examined across three contributions, none were found to clearly refute any component. The RAVENEA benchmark itself (10 candidates examined, 0 refutable) appears novel within this limited search scope, as does the Culture-Aware Contrastive learning framework (10 candidates, 0 refutable) and RegionScore metric (10 candidates, 0 refutable). The absence of refutable prior work across all contributions, combined with the sparse benchmark leaf containing only one sibling paper, suggests these contributions occupy relatively unexplored territory. However, this assessment is constrained by the top-30 semantic search scope and does not constitute exhaustive coverage of all potentially relevant prior work.

Based on the limited literature search, RAVENEA appears to introduce distinct evaluation infrastructure for a nascent research direction. The sparse benchmark category and zero refutable candidates across contributions indicate novelty within the examined scope, though the small search scale (30 papers) means potentially relevant work outside top semantic matches may exist. The taxonomy structure confirms that dedicated benchmarks for retrieval-augmented cultural understanding remain rare compared to broader heritage or RAG research.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: RAVENEA benchmark for multimodal retrieval-augmented visual culture understanding

Description: The authors introduce RAVENEA, the first benchmark explicitly designed to evaluate vision-language models in using external knowledge for visual culture understanding. It covers eight countries across eleven categories, linking images to human-ranked Wikipedia documents on two tasks: culture-focused visual question answering and culture-informed image captioning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Event-Enriched Image Analysis Grand Challenge At ACM Multimedia 2025

URL: View paper

Brief Assessment

Event Image Challenge[16] focuses on event-level understanding in news/journalism contexts (who, when, where, what, why of events), while RAVENEA targets cultural understanding across countries with retrieval-augmented generation for visual culture tasks. These are distinct application domains and objectives.

---

2. Multi-Modal Semantic Parsing for the Interpretation of Tombstone Inscriptions

URL: View paper

Brief Assessment

Tombstone Parsing[14] focuses on heritage preservation through tombstone digitization using VLMs and RAG for structured semantic parsing. This is a domain-specific application (tombstone interpretation) rather than a general benchmark for evaluating vision-language models on visual culture understanding across multiple countries and tasks.

---

3. LUMOS-DM: Landscape-Based Multimodal Scene Retrieval Enhanced by Diffusion Model

URL: View paper

Brief Assessment

LUMOS-DM[33] focuses on landscape-based multimodal scene retrieval using diffusion models for video content retrieval. It does not address visual culture understanding, culturally-grounded benchmarks, or retrieval-augmented generation for cultural context in vision-language models.

---

4. ArtRAG: Retrieval-Augmented Generation with Structured Context for Visual Art Understanding

URL: View paper

Brief Assessment

ArtRAG[2] focuses specifically on visual art understanding using knowledge graphs for artwork explanation, while the original paper addresses broader visual culture understanding across multiple countries and categories using Wikipedia documents. The domains and methodologies are distinct.

---

5. GREEN: Generative Retrieval-Enhanced Emotional Support Conversations

URL: View paper

Brief Assessment

GREEN[62] focuses on emotional support conversations with retrieval-enhanced generation, not on visual culture understanding benchmarks for vision-language models. The candidate addresses a completely different domain (conversational emotional support) than the original paper's multimodal cultural understanding evaluation.

---

6. From local concepts to universals: Evaluating the multicultural understanding of vision-language models

URL: View paper

Brief Assessment

Multicultural VLM Understanding[61] focuses on evaluating VLMs' inherent multicultural understanding through retrieval and grounding tasks without external knowledge augmentation, whereas RAVENEA specifically evaluates retrieval-augmented generation with human-ranked Wikipedia documents for cultural understanding.

---

7. Exposing Blindspots: Cultural Bias Evaluation in Generative Image Models

URL: View paper

Brief Assessment

Cultural Bias Images[29] focuses on evaluating cultural bias in generative image models (text-to-image and image-to-image) rather than retrieval-augmented visual culture understanding benchmarks for vision-language models. The candidate does not address retrieval-augmented generation or multimodal retrieval systems.

---

8. Lost in Translation: A Position Paper on Probing Cultural Bias in Vision-Language Models via Hanbok VQA

URL: View paper

Brief Assessment

Hanbok VQA[63] focuses on evaluating cultural bias in VLMs specifically for Korean traditional attire through a hierarchical VQA task, not on building a general multimodal retrieval-augmented benchmark for visual culture understanding across multiple countries and categories.

---

9. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration

URL: View paper

Brief Assessment

k-viscuit[60] focuses on evaluating Korean cultural understanding through visual recognition and reasoning tasks using a semi-automated human-VLM collaborative framework. It does not address retrieval-augmented generation or multimodal retrieval systems, which are central to RAVENEA's contribution.

---

10. Cultural Heritage Assistant: A Lightweight Retrieval Augmented Generation Method Enhanced Vision-Language Model for Cultural Heritage

URL: View paper

Brief Assessment

Heritage Assistant RAG[39] focuses on cultural heritage artifact question-answering using a lightweight RAG method for VLMs, not on creating a benchmark for evaluating multimodal retrieval-augmented visual culture understanding across multiple countries and categories.

Contribution 2: Culture-Aware Contrastive (CAC) learning framework

Description: The authors propose Culture-Aware Contrastive learning, a supervised learning framework that enhances cultural awareness in multimodal retrieval by incorporating culture-targeted annotations. This framework is compatible with both CLIP and SigLIP architectures and demonstrates marked gains in retrieval accuracy.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Multimodal cultural safety: Evaluation frameworks and alignment strategies

URL: View paper

Brief Assessment

Multimodal Cultural Safety[41] focuses on cultural safety evaluation and alignment strategies for vision-language models, not on contrastive learning frameworks for cultural awareness in multimodal retrieval.

---

2. â¦ Manifold Disentanglement Transformer (IMD-Transformer): A Robust Framework for Low-Resource Cross-Modal Learning in Digital Cultural Heritage and Tourism

URL: View paper

Brief Assessment

IMD-Transformer[48] focuses on manifold disentanglement for cross-modal learning in cultural heritage contexts, while the original paper proposes a supervised contrastive framework specifically for cultural relevance classification in multimodal retrieval with Wikipedia documents.

---

3. Computational Approaches to Cross-Cultural Multimodal Film and TV Integration Using AMTAs

URL: View paper

Brief Assessment

Cross-Cultural Film AMTAs[47] focuses on machine translation for film/TV data using variational inference and contrastive learning for small-sample optimization, not on multimodal retrieval with culture-targeted annotations for vision-language models.

---

4. AraTraditions10k bridging cultures with a comprehensive dataset for enhanced cross lingual image annotation retrieval and tagging

URL: View paper

Brief Assessment

AraTraditions10k[13] focuses on cross-lingual image annotation and retrieval for Arabic-English datasets using attention mechanisms and contrastive loss in W2VV models, not on culture-aware contrastive learning frameworks for multimodal retrieval as proposed in the original paper.

---

5. CultureCLIP: Empowering CLIP with Cultural Awareness through Synthetic Images and Contextualized Captions

URL: View paper

Brief Assessment

CultureCLIP[44] focuses on enhancing CLIP with cultural awareness through synthetic images and contextualized captions for fine-grained concept recognition. The original paper's CAC framework is designed for multimodal retrieval with culture-targeted annotations to retrieve culturally relevant documents, which is a different application domain and methodology.

---

6. Cultural bias mitigation in vision-language models for digital heritage documentation: A comparative analysis of debiasing techniques

URL: View paper

Brief Assessment

Heritage Debiasing[43] focuses on cultural bias mitigation in heritage documentation using cross-modal adapters for debiasing, not on culture-aware contrastive learning for multimodal retrieval. The technical objectives and application domains differ fundamentally.

---

7. No filter: Cultural and socioeconomic diversity in contrastive vision-language models

URL: View paper

Brief Assessment

No Filter[45] focuses on training data composition (English-only vs. global multilingual data) for contrastive VLMs like SigLIP, not on supervised culture-targeted annotation frameworks. The original paper's CAC uses explicit cultural relevance labels for retrieval, which is architecturally distinct from No Filter's data filtering approach.

---

8. Matina: A culturally-aligned Persian language model using multiple LoRA experts

URL: View paper

Brief Assessment

Matina[42] focuses on Persian language model alignment using multiple LoRA experts across specific cultural domains (culinary, tourism, socio-culture), not on developing a contrastive learning framework for multimodal retrieval with culture-targeted annotations as in the original paper.

---

9. ExplainHM++: Explainable Harmful Meme Detection With Retrieval-Augmented Debate Between Large Multimodal Models

URL: View paper

Brief Assessment

ExplainHM++[49] focuses on harmful meme detection with retrieval-augmented debate between multimodal models, not on cultural awareness in multimodal retrieval. The candidate's minimal context does not demonstrate prior work on culture-aware contrastive learning frameworks.

---

10. Finding Culture-Sensitive Neurons in Vision-Language Models

URL: View paper

Brief Assessment

Culture-Sensitive Neurons[46] focuses on identifying and ablating culture-sensitive neurons within existing VLMs to understand internal representations, rather than proposing a new contrastive learning framework for training multimodal retrievers with cultural annotations.

Contribution 3: RegionScore metric for evaluating cultural relevance in image captions

Description: The authors introduce RegionScore, a novel evaluation metric that quantifies the extent to which generated captions reference specific geopolitical regions. This metric addresses the mismatch between automatic metrics and human judgments of cultural appropriateness in image captioning tasks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. On the cultural gap in text-to-image generation

URL: View paper

Brief Assessment

Cultural Gap Generation[54] focuses on text-to-image generation quality across cultures, not image captioning evaluation. Their work addresses filtering training data for T2I models using multi-modal alignment metrics, which is a different task from evaluating cultural relevance in generated captions.

---

2. Understanding and evaluating racial biases in image captioning

URL: View paper

Brief Assessment

Racial Biases Captioning[51] focuses on racial bias in image captioning using sentiment analysis and vocabulary differences, not on cultural relevance metrics for geopolitical regions.

---

3. Vision-language models under cultural and inclusive considerations

URL: View paper

Brief Assessment

Cultural VLM Considerations[56] does not propose RegionScore or any similar metric for quantifying geopolitical region references in captions. The candidate focuses on evaluating existing VLMs using standard metrics (BLEU, METEOR, CIDEr, SPICE) and human evaluation for cultural image captioning, without introducing novel evaluation metrics.

---

4. Cic: A framework for culturally-aware image captioning

URL: View paper

Brief Assessment

CIC Framework[52] does not propose RegionScore or any similar metric for evaluating cultural relevance. Instead, it focuses on generating culturally-aware captions through VQA and LLM prompting, and introduces Culture Noise Rate (CNR) as an evaluation metric, which measures the ratio of cultural words rather than geopolitical region references.

---

5. Beyond words: Exploring cultural value sensitivity in multimodal models

URL: View paper

Brief Assessment

Cultural Value Sensitivity[50] focuses on evaluating cultural value alignment in multimodal models using World Values Survey questions and does not propose evaluation metrics for image captioning tasks. The papers address different aspects of cultural understanding in multimodal systems.

---

6. Beyond aesthetics: Cultural competence in text-to-image models

URL: View paper

Brief Assessment

Cultural Competence Models[55] focuses on evaluating cultural competence in text-to-image generation models, not image captioning. The candidate does not address evaluation metrics for cultural relevance in image captions.

---

7. Quantifying and Mitigating Dataset Biases in Video Under-standing Tasks across Cultural Contexts

URL: View paper

Brief Assessment

Cultural Video Biases[59] focuses on quantifying dataset biases in video understanding tasks, while the original paper introduces RegionScore specifically for evaluating geopolitical region references in image captions. The candidate's limited context does not provide sufficient detail about their metric design to establish prior work on region-specific caption evaluation.

---

8. Culturally-aware Image Captioning

URL: View paper

Brief Assessment

Culturally-aware Captioning[53] focuses on generating culturally-aware captions and mentions the need for new evaluation metrics, but does not present a specific metric like RegionScore. The original paper introduces RegionScore as a concrete implementation to quantify geopolitical region references in captions.

---

9. Semantic and Expressive Variations in Image Captions Across Languages

URL: View paper

Brief Assessment

Semantic Caption Variations[58] focuses on semantic and expressive variation across languages in image captions, not on evaluating cultural relevance. Their metrics measure linguistic diversity and scene graph coverage, not geopolitical region references.

---

10. How Culturally Aware Are Vision-Language Models?

URL: View paper

Brief Assessment

Cultural VLM Awareness[57] introduces the Cultural Awareness Score (CAS), a binary metric measuring presence/absence of culturally-specific information in captions. This differs from RegionScore, which quantifies geopolitical region references in captions using a continuous score based on country names and demonyms.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] RAVENEA: A Benchmark for Multimodal Retrieval-Augmented Visual Culture Understanding View paper
- [1] Retrieval augmented generation and understanding in vision: A survey and new outlook View paper
- [2] ArtRAG: Retrieval-Augmented Generation with Structured Context for Visual Art Understanding View paper
- [3] DIR: Retrieval-Augmented Image Captioning with Comprehensive Understanding View paper
- [4] FolkRAG: a retrieval-augmented generation system for cultural heritage materials View paper
- [5] Valuesrag: Enhancing cultural alignment through retrieval-augmented contextual learning View paper
- [6] Leveraging Retrieval-Augmented Generation for Culturally Inclusive Hakka Chatbots: Design Insights and User Perceptions View paper
- [7] Multimedia information retrieval using content-based image retrieval and context link for Chinese cultural artifacts View paper
- [8] Probing historical image contexts: Enhancing visual archive retrieval through computer vision View paper
- [9] Video-RAG: Visually-aligned Retrieval-Augmented Long Video Comprehension View paper
- [10] RegionRAG: Region-level Retrieval-Augmented Generation for Visual Document Understanding View paper
- [11] Evcap: Retrieval-Augmented Image Captioning with External Visual-Name Memory for Open-World Comprehension View paper
- [12] Towards Cross-Modal Retrieval in Chinese Cultural Heritage Documents: Dataset and Solution View paper
- [13] AraTraditions10k bridging cultures with a comprehensive dataset for enhanced cross lingual image annotation retrieval and tagging View paper
- [14] Multi-Modal Semantic Parsing for the Interpretation of Tombstone Inscriptions View paper
- [15] CAIRe: Cultural Attribution of Images by Retrieval-Augmented Evaluation View paper
- [16] Event-Enriched Image Analysis Grand Challenge At ACM Multimedia 2025 View paper
- [17] Immersive Virtual Museums with Spatially-Aware Retrieval-Augmented Generation View paper
- [18] Computer vision and AI TOOLS for enhancing user experience in the cultural heritage domain View paper
- [19] Diffusion Based Augmentation for Captioning and Retrieval in Cultural Heritage View paper
- [20] How Can Generative Artificial Intelligence Techniques Facilitate Intelligent Research into Ancient Books? View paper
- [21] Cultural heritage information retrieval: past, present, and future trends View paper
- [22] Ontology-based Secure Retrieval of Semantically Significant Visual Contents View paper
- [23] Evaluating cultural knowledge processing in large language models: a cognitive benchmarking framework integrating retrieval-augmented generation View paper
- [24] When MLLMs Meet ICH: A visual retrieval-augmented generation-based method for intangible cultural heritage image recognition-take Shadow Puppetry as a case View paper
- [25] Mosaic: Finding artistic connections across culture with conditional image retrieval View paper

- [26] Vision-Augmented RAG System for Interactive Local Heritage Exploration View paper
- [27] Signs as Tokens: A Retrieval-Enhanced Multilingual Sign Language Generator View paper
- [28] FITMag: A Framework for Generating Fashion Journalism Using Multimodal LLMs, Social Media Influence, and Graph RAG View paper
- [29] Exposing Blindspots: Cultural Bias Evaluation in Generative Image Models View paper
- [30] Content-based image retrieval of cultural heritage symbols by interaction of visual perspectives View paper
- [31] A review of image retrieval methods for digital cultural heritage resources View paper
- [32] Visual information retrieval from historical document images View paper
- [33] LUMOS-DM: Landscape-Based Multimodal Scene Retrieval Enhanced by Diffusion Model View paper
- [34] A hybrid ontology and visual-based retrieval model for cultural heritage multimedia collections View paper
- [35] Managing digital cultural objects: analysis, discovery and retrieval View paper
- [36] Exploring the Integration of National Cultural Resources through Content Retrieval and Visual Semantic Segmentation Display View paper
- [37] Content-Based Indexing and Retrieval of Cultural Heritage Data: An Integrated Approach to Documentation with Application to the EROS View paper
- [38] Leading to Cultural Resources Through Visuals: An Image Retrieval System and On-site Information Approach View paper
- [39] Cultural Heritage Assistant: A Lightweight Retrieval Augmented Generation Method Enhanced Vision-Language Model for Cultural Heritage View paper
- [40] A Network Perspective on Gradient Flow Equations for Deep Linear Neural Networks View paper
- [41] Multimodal cultural safety: Evaluation frameworks and alignment strategies View paper
- [42] Matina: A culturally-aligned Persian language model using multiple LoRA experts View paper
- [43] Cultural bias mitigation in vision-language models for digital heritage documentation: A comparative analysis of debiasing techniques View paper
- [44] CultureCLIP: Empowering CLIP with Cultural Awareness through Synthetic Images and Contextualized Captions View paper
- [45] No filter: Cultural and socioeconomic diversity in contrastive vision-language models View paper
- [46] Finding Culture-Sensitive Neurons in Vision-Language Models View paper
- [47] Computational Approaches to Cross-Cultural Multimodal Film and TV Integration Using AMTAs View paper
- [48] â¦ Manifold Disentanglement Transformer (IMD-Transformer): A Robust Framework for Low-Resource Cross-Modal Learning in Digital Cultural Heritage and Tourism View paper
- [49] ExplainHM++: Explainable Harmful Meme Detection With Retrieval-Augmented Debate Between Large Multimodal Models View paper
- [50] Beyond words: Exploring cultural value sensitivity in multimodal models View paper
- [51] Understanding and evaluating racial biases in image captioning View paper
- [52] Cic: A framework for culturally-aware image captioning View paper
- [53] Culturally-aware Image Captioning View paper
- [54] On the cultural gap in text-to-image generation View paper
- [55] Beyond aesthetics: Cultural competence in text-to-image models View paper
- [56] Vision-language models under cultural and inclusive considerations View paper
- [57] How Culturally Aware Are Vision-Language Models? View paper
- [58] Semantic and Expressive Variations in Image Captions Across Languages View paper
- [59] Quantifying and Mitigating Dataset Biases in Video Under-standing Tasks across Cultural Contexts View paper
- [60] Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration View paper
- [61] From local concepts to universals: Evaluating the multicultural understanding of vision-language models View paper
- [62] GREEN: Generative Retrieval-Enhanced Emotional Support Conversations View paper
- [63] Lost in Translation: A Position Paper on Probing Cultural Bias in Vision-Language Models via Hanbok VQA View paper