Multimodal Boolean Logic Culture Retrieval

Jian Zhang[1], Shuheng Hu[1], Jiutian Chang[1], Junyi Guo[1], Huanda Lu[2], Qiufeng Wang[1], Fangyu Wu[1,*]

[1]Xi'an Jiaotong-Liverpool University, Suzhou, China, [2]NingboTech University, Ningbo, China

Correspondence: Fangyu.Wu02@xjtlu.edu.cn

Abstract

Cultural image retrieval is essential for organizing and analyzing large-scale digital heritage resources. However, existing multimodal retrieval methods usually encode a query as a holistic embedding, making it difficult to preserve logical operators and cultural relations during retrieval. Therefore, we introduce Multimodal Boolean Logic Culture Retrieval, a novel task that moves beyond general visual similarity by enforcing strict Boolean constraints over structured cultural elements. To tackle this, we propose CLAIR, a training-free framework that reformulates the retrieval objective to seamlessly integrate logical operations directly within the multimodal embedding space. Furthermore, we construct CulTi2.0, a dedicated benchmark designed to evaluate diverse logical query forms based on cultural visual atoms. Extensive experiments show that while current multimodal models struggle with explicit logical queries, CLAIR provides consistent and generalizable improvements as a plug-and-play solution across various datasets and model families. Code and dataset are publicly available at https://github.com/AI4Culture/CulTi2.0.

1 Introduction

Cultural heritage retrieval requires a fundamentally different formulation from general-purpose image search (Tsai, 2007). Rather than using free-form natural language, cultural knowledge is structured through controlled entities and attributes (Carriero et al., 2019; Fan and Wang, 2022). A ritual bronze vessel, for instance, is retrieved not by a holistic caption, but by specific logical constraints over these visual atoms, such as "bronze vessel" AND "dragon motif," or "bronze vessel" NOT "bird pattern." Motivated by this observation, we introduce **Multimodal Boolean Logic Culture Retrieval**, a new task that formalizes cultural image search as explicit Boolean compositions over typed visual concepts.

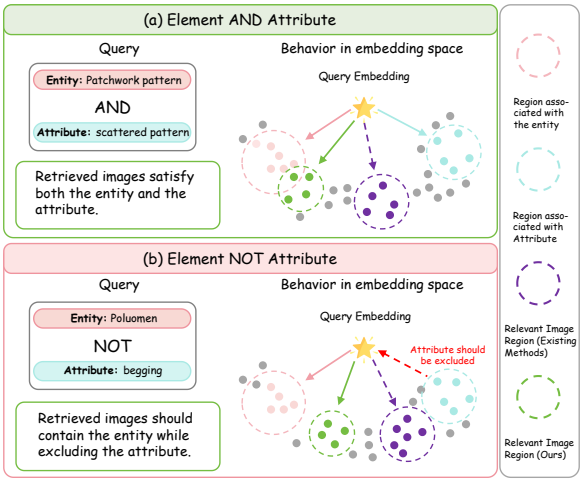


Figure 1: Conceptual illustration of Multimodal Boolean Logic Culture Retrieval. The purple regions denote the retrieval space of existing methods relying on semantic overlap. The green regions indicate the retrieval space of our method satisfying explicit Boolean constraints.

However, realizing this task faces two fundamental challenges: (1) **Structural Limitations of Multimodal Models.** Prevailing models like CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) encode queries into holistic embeddings that inherently collapse complex logical structures. Entities and modifiers are treated as independent signals rather than bound relations, and negation becomes a weak penalty rather than a hard constraint. Consequently, logically excluded but semantically related concepts often rank incorrectly high (Park et al., 2025; Jha et al., 2026; van den Elsen et al., 2025). (2) **Absence of Dedicated Benchmarks.** No existing benchmark evaluates retrieval under explicit logical constraints for cultural visual concepts. This absence conflates logical comprehen-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37380–37389
October 24-29, 2026 ©2026 Association for Computational Linguistics

sion with mere cultural relevance, hindering the principled diagnosis of model failure modes.

To overcome the structural limitations of existing embeddings, we propose CLAIR (**C**ulture **L**ogic-**A**ware **I**mage **R**etrieval), a training-free framework. The core idea is to reformulate the retrieval objective using two complementary forms of evidence: modification and logical relations. First, we introduce an embedding-level attention consistency score that encourages the retrieval function to jointly activate the modifier, the target entity, and their composed expression. Second, we adapt fuzzy-logic operations (Klir and Yuan, 1995) to transform discrete Boolean operators into continuous score-level computations. As shown in Fig. 1, although CLAIR operates entirely over frozen embeddings, it makes the retrieval highly sensitive to the logical dependencies among cultural elements, effectively enforcing hard constraints.

To address the absence of evaluation resources for this task, we build CulTi2.0 on the cultural heritage dataset CulTi (Yuan et al., 2025) and extend it into a dedicated testbed for cultural Boolean retrieval. Existing datasets mainly test free-form descriptive matching, where relevance is determined by holistic semantic overlap. In contrast, CulTi2.0 strictly organizes searches around typed visual atoms. By evaluating four explicit Boolean forms (*Entity AND Entity*, *Entity AND Attribute*, *Entity NOT Entity*, and *Entity NOT Attribute*), it allows us to test whether a model truly satisfies the intended logical constraints rather than merely retrieving culturally related images. This design turns cultural image search into a rigorously controlled environment, decoupling visual recognition from logical comprehension.

We conduct extensive experiments across three benchmarks, including our proposed CulTi2.0 and two general-domain negation benchmarks (Park et al., 2025; Parcalabescu et al., 2022), with a broad set of open-source and closed-source multimodal embedding models. The limited performance of existing models on CulTi2.0 reveals that Multimodal Boolean Logic Culture Retrieval remains a challenging and underexplored task. With its training-free design, CLAIR can be applied as a plug-and-play scoring function across different embedding models, consistently improving performance across query types involving both modifier relations and logical relations. These results demonstrate the effectiveness of our framework and highlight Multimodal Boolean Logic Culture Retrieval as a new perspective for multimodal retrieval, shifting the focus from visual similarity to logical and relational relevance.

In summary, our main contributions are as follows: (1) We introduce Multimodal Boolean Logic Culture Retrieval, a novel task that moves beyond visual similarity by requiring models to satisfy explicit modification and logical relations. (2) We propose CLAIR, a training-free logic-aware framework that reformulates the retrieval objective by integrating semantic, modification, and logical evidence within the embedding space. (3) We construct CulTi2.0, a dedicated benchmark evaluating this task through combinations of cultural concepts, attributes, and Boolean operators. (4) Extensive experiments reveal that while existing multimodal models struggle with explicit logical queries, CLAIR provides consistent and generalizable improvements across various datasets and model families.

2 Related Work

2.1 Multimodal Cultural Retrieval

Multimodal retrieval has mainly followed two paradigms. The first is the dual-encoder paradigm, where models such as CLIP, BLIP, and SigLIP families learn aligned image–text representations through contrastive pretraining (Radford et al., 2021; Li et al., 2022; Zhai et al., 2023; Tschannen et al., 2025). The second is multimodal embedding models, such as Qwen3-VL-Emb (Li et al., 2026), further improve cross-modal alignment, visual representation, and multilingual retrieval.

Recent retrieval works have started focusing on cultural domains. Cultural retrieval benchmarks such as CulTi (Yuan et al., 2025) focus on cross-modal retrieval for Chinese cultural heritage documents, whereas C^3 (Zhang et al., 2025) improves cultural retrieval by augmenting completeness and consistency with LLM-driven methods. Broader cultural understanding benchmarks, such as TimeTravel (Ghaboura et al., 2025) and RAVENEA (Li et al., 2025), further evaluate multimodal models for retrieving cultural context. These benchmarks emphasize descriptive or contextual relevance rather than explicit Boolean constraints over cultural entities and attributes.

2.2 Boolean Logic in Retrieval

Boolean logic provides a natural way to express structured retrieval conditions such as conjunction,

disjunction, and exclusion, but these operators are difficult to enforce in neural retrieval because dense models rank continuous query embeddings by similarity (Karpukhin et al., 2020; Izacard et al., 2021; Xiong et al., 2021). Recent work has studied this issue in text retrieval through benchmarks such as NevIR and BoolQuestions (van den Elsen et al., 2025; Zhang et al., 2024), and in multimodal retrieval through compositional or exclusion-based settings such as MA-CIR and ExclMM (Byun et al., 2025; Jha et al., 2026). Methodologically, ConE and BetaE model logical operations in specialized embedding spaces (Zhang et al., 2021; Ren and Leskovec, 2020), ExclMM trains additional disentanglement modules (Jha et al., 2026), while NSFL and DEO explore training-free correction over frozen embeddings (Vexler et al., 2026; Lee et al., 2026). However, prior work has not examined Boolean retrieval in cultural settings, where entities and attributes are distinct semantic types but often strongly correlated. We address this gap with a training-free framework that jointly models semantic relevance, entity-attribute binding, and logical validity over frozen multimodal embeddings.

3 Method

To address Multimodal Boolean Logic Culture Retrieval, we propose a training-free framework, CLAIR, that follows two design principles. First, the vision-language encoder is kept frozen, so the method can be directly applied to existing embedding-based retrieval systems without model-specific training. Second, it supplements the standard retrieval objective with fine-grained logical evidence computed in the embedding space, enabling retrieval to reflect logical constraints in culture domain. The overview of CLAIR is shown in Fig. 2.

3.1 Query Parsing

Existing retrieval models treat all query tokens as part of one semantic representation. In Multimodal Boolean Logic Culture Retrieval, however, different tokens play different relational roles. Therefore, we introduce a lightweight query parsing step that converts a cultural query into multiple components. Given a query q, CLAIR first parses it into multiple components. This step separates the target of retrieval from the cultural specification and logical constraint of the query, providing structured inputs for subsequent relation-aware scoring.

Specifically, we parse a query into three components: cultural concepts, cultural modifiers, and logical operators. A cultural concept refers to a target entity, such as *silk* or *vase*. A cultural modifier refers to an attribute, pattern, or symbolic feature, such as *Dunhuang* or *blue-and-white*. Logical operators include *AND*, *OR*, and *NOT*. For example, the query *Dunhuang AND silk* contains the modifier *Dunhuang*, the concept *silk*, and the logical operator *AND*. Besides their co-occurrence, we further interpret this query as a modifier relation, where the modifier should be associated with the target concept. In contrast, the query *silk NOT Dunhuang* requires the image to match *silk* while suppressing the *Dunhuang* modifier.

3.2 Semantic Scores

Following standard embedding-based image retrieval, we first compute semantic similarity scores $S_{\text{semantic}}(I, q)$ using a frozen vision-language encoder for each parsed component, where I denotes the image. We denote the normalized similarity score between image I and a text component x as $s_x(I)$, where x denotes a parsed text component. The score is normalized to $[0, 1]$ for subsequent computation. This step preserves the global semantic relevance between the image and the original query as the basic retrieval signal.

3.3 Modifier Relation Score

Existing retrieval often treats a cultural modifier and a target concept as independent cues, thereby failing to distinguish true modifier relations from co-occurrence. We address this by introducing an embedding-level attention consistency score. It encourages the image to provide balanced support for the modifier and the concept, while the composed phrase score further verifies whether they form a meaningful cultural composition.

The modifier relation score aims to distinguish a modification relation from simple co-occurrence. For a modifier m and a concept o, we construct a composed phrase p, such as *Dunhuang silk*. We then obtain three normalized semantic scores: $s_m(I)$ for the modifier, $s_o(I)$ for the concept, and $s_p(I)$ for the composed phrase.

We then use an attention mechanism to estimate whether the image includes the modifier and the concept. The attention weights are computed as:

$$a_m(I) = \frac{\exp(s_m(I)/\tau)}{\exp(s_m(I)/\tau) + \exp(s_o(I)/\tau)}, \quad (1)$$

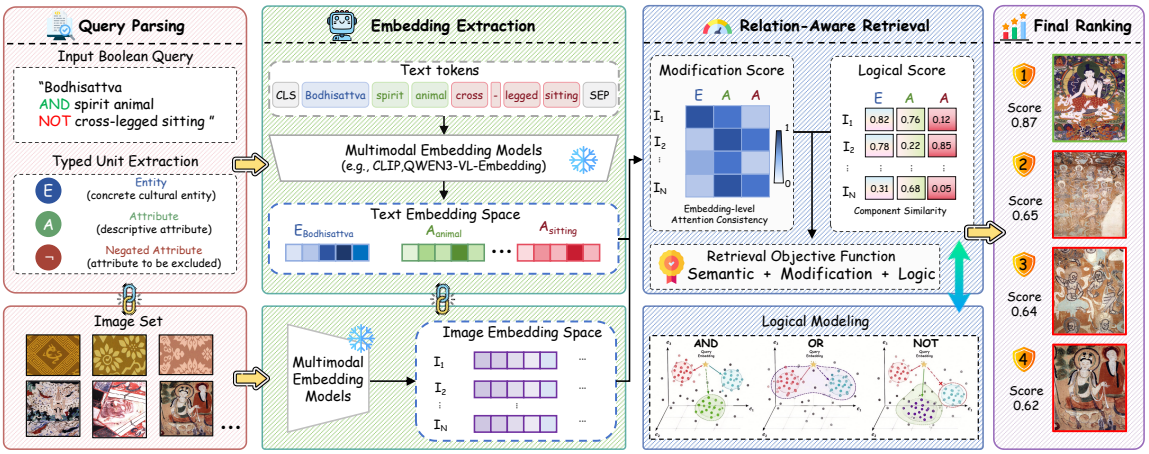


Figure 2: Overview of CLAIR. The framework parses an input Boolean query into typed units, extracts text and image embeddings from frozen multimodal encoders, computes relation-aware retrieval scores by combining semantic, modification, and logical evidence, and produces the final ranking.

$$a_o(I) = \frac{\exp(s_o(I)/\tau)}{\exp(s_m(I)/\tau) + \exp(s_o(I)/\tau)}, \quad (2)$$

where τ is a temperature parameter. If an image mainly matches the modifier but not the concept, or vice versa, the attention distribution becomes imbalanced. Therefore, we define the attention consistency as:

$$C_{\text{att}}(m, o, I) = 4a_m(I)a_o(I). \quad (3)$$

This term reaches its maximum when the image provides balanced semantic support for both attribute and concept. To further encourage a meaningful composition of modifier-concept, we also include the composition score of the phrase $s_p(I)$.

The modifier relation score is defined as:

$$S_{\text{modifier}}(m, o, I) = s_p(I) \cdot \sqrt{s_m(I)s_o(I)} \cdot C_{\text{att}}(m, o, I). \quad (4)$$

This score becomes high only when the image matches the modifier, the concept, and their composed expression simultaneously. For example, an image should receive a high score for *Dunhuang silk* only when it is semantically close to both *Dunhuang* and *silk*, as well as the composed phrase.

If a query contains multiple modifier-concept relations, we aggregate their scores by averaging:

$$S_{\text{modifier}}(I, q) = \frac{1}{K} \sum_{k=1}^{K} S_{\text{modifier}}(m_k, o_k, I), \quad (5)$$

where K denotes the number of modifier-concept relations in the query.

3.4 Logical Relation Score

The logical relation score is designed to prevent logical operators from being absorbed as ordinary words in a holistic query embedding. In contrast, we apply lightweight fuzzy-logic operations to explicitly model conjunction, disjunction, and exclusion. In this way, *AND* requires multiple components to be jointly supported, *OR* allows either component to provide evidence, and *NOT* suppresses images that match the excluded cultural cue.

AND. For an AND query, the image should satisfy both components. Given two component scores $s_X(I)$ and $s_Y(I)$, we use conjunction:

$$S^{\text{AND}}_{\text{logic}}(I, q) = \sqrt{s_X(I)s_Y(I)}. \quad (6)$$

For example, in *Dunhuang AND silk*, the score is high only when the image is semantically close to both *Dunhuang* and *silk*.

OR. For an OR query, the image only needs to satisfy one of the components. We use disjunction:

$$S^{\text{OR}}_{\text{logic}}(I, q) = 1 - (1 - s_X(I))(1 - s_Y(I)). \quad (7)$$

This score becomes high if either component is strongly supported by the image.

NOT. For a NOT query, the image should match the positive component while suppressing the negative component. We define negation:

$$S^{\text{NOT}}_{\text{logic}}(I, q) = s_X(I)(1 - s_Y(I)). \quad (8)$$

For example, in *silk NOT Dunhuang*, the score is high only when the image matches *silk* and has low similarity to *Dunhuang*. The final logic score $S_{\text{logic}}(I, q)$ is selected according to the logical operator parsed from the query.

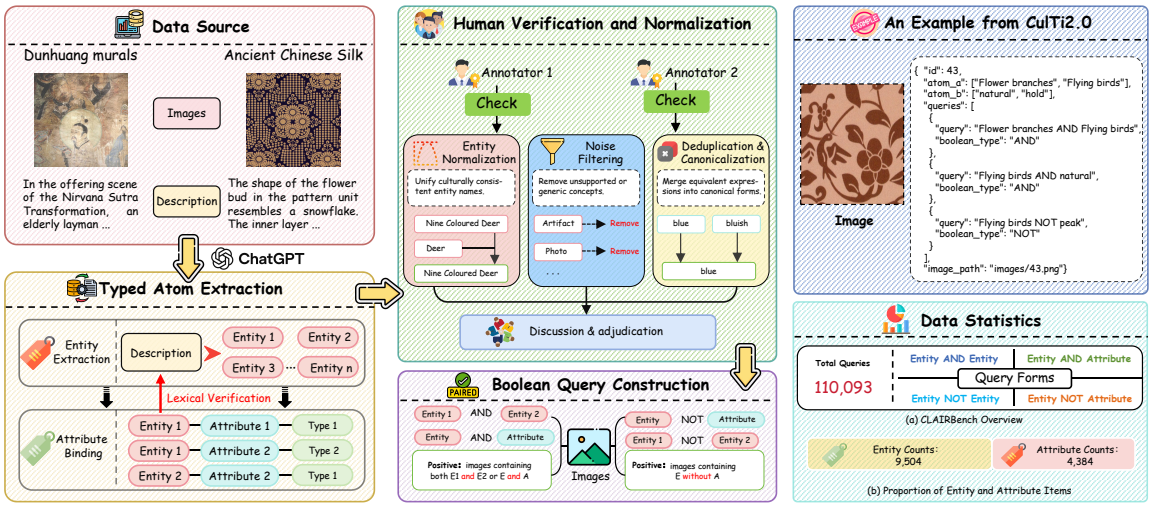


Figure 3: Overview of CulTi2.0 construction. We derive typed entity–attribute atoms from culturally grounded image descriptions, verify and normalize them through human annotation, and compose four Boolean query types. The right panel shows an example data instance and dataset statistics.

3.5 Retrieval Function

The final retrieval function combines global semantic matching, modifier relation satisfaction, and logical satisfaction as:

$$S_{\text{final}}(I,q) = \alpha S_{\text{semantic}}(I,q) + \beta S_{\text{modifier}}(I,q) + \gamma S_{\text{logic}}(I,q). \quad (9)$$

This formulation changes the ranking criterion from similarity-only matching to relation-aware matching. Such a shift follows directly from the goal of the Multimodal Boolean Logic Culture Retrieval task, because a correct result should not merely contain semantically related visual cues but also satisfy the cultural relations specified by the query. Accordingly, the global term preserves basic query relevance, while the modifier and logic terms encode the two task-specific criteria. Thus, a high score reflects both semantic content matching and relational intent matching.

4 CulTi2.0 Benchmark

We propose CulTi2.0 to evaluate whether multimodal embeddings can satisfy Boolean constraints over fine-grained cultural concepts. By representing retrieval intents through culturally grounded entities, visually verifiable attributes, and logical operators, CulTi2.0 strictly separates mere semantic relevance from logical validity.

4.1 Dataset Construction.

Data Source. We build CulTi2.0 upon CulTi (Yuan et al., 2025), a curated dataset of Chinese cultural heritage with 5,726 image–text pairs spanning ancient Chinese silk artifacts and Dunhuang murals. CulTi provides fine-grained, expert-aligned descriptions with dense co-occurrences of cultural elements and attributes, making it suitable for constructing non-trivial Boolean queries.

Typed Atom Extraction. We decompose each natural description into typed atoms through a strict two-stage extraction and verification process. First, we use ChatGPT-5.2 (Singh et al., 2025) to identify candidate entities. Second, we bind each entity to its modifying attributes and assign each attribute to a predefined type (e.g., color, posture, material), producing tuples of the form (e, a, t). To ensure grounding, we apply lexical verification: every extracted entity and attribute must trace back to the source text through exact matching, alias matching, or manually defined normalization rules. Tuples containing unsupported atoms are discarded, effectively preventing hallucinated concepts from entering the benchmark.

Human Verification and Normalization. After lexical grounding, two independent annotators with cultural heritage domain expertise manually curate the candidate tuples to resolve structural issues that automatic extraction cannot reliably handle. Specifically, they address generic concepts, entity granularity, and entity–attribute alignment. For instance, when multiple granularities exist, we retain the culturally specific expression (e.g., keeping *Nine-Colored Deer* rather than a generic *deer*). Furthermore, they normalize equivalent expressions into canonical forms to avoid duplicate queries. Disagreements are resolved through discussion, ul-

timately yielding a high-quality, curated pool of description-supported entities and attributes.

Boolean Query Construction. We construct Boolean queries exclusively from the verified atoms, ensuring that every query is traced back to explicit textual evidence. The construction produces four query forms. *Entity AND Entity* queries are generated from entities co-occurring in the same description. *Entity AND Attribute* queries are derived directly from verified (e, a, t) tuples. For *Entity NOT Entity*, we pair a verified entity with a mutually exclusive entity sampled from the same cultural category. For *Entity NOT Attribute*, we use a conservative counter-attribute strategy: rather than treating an unmentioned attribute as absent, we take a verified tuple (e, a^+, t) and select a mutually exclusive attribute a^- from the same type t to construct e NOT a^-. By anchoring negations in positive evidence of incompatibility, we rigorously evaluate whether a retriever follows the intended Boolean operator rather than merely matching mentioned concepts.

4.2 Dataset Statistics

The final CulTi2.0 benchmark contains 110,093 Boolean query pairs, including 94,293 *Entity AND Entity*, 6,511 *Entity AND Attribute*, 5,109 *Entity NOT Entity*, and 4,180 *Entity NOT Attribute* queries. We keep the original CulTi (Yuan et al., 2025) splits and construct queries separately within each split to avoid data leakage. Dataset examples and statistics are provided in Fig. 3.

5 Experiments

5.1 Experimental Setup

Implementation Details. For open-source models, we compute embeddings on an NVIDIA A6000 GPU. For closed-source models, we obtain multimodal embeddings through official APIs. We report Recall@1, Recall@5, Recall@10, and RSum for both text-to-image and image-to-text retrieval. For multiple-choice evaluation, we select the candidate with the highest score and report accuracy.

Models. We evaluate both open-source and closed-source multimodal embedding models. The open-source models include ALBEF (Li et al., 2021), BLIP (Li et al., 2022), CN-CLIP (Yang et al., 2022), SigLIP2 (Tschannen et al., 2025), and Qwen3-VL-Emb (Li et al., 2026). The closed-source models include Tongyi-Flash, Tongyi-Plus, and Multimodal-Embedding-v1, corresponding to the official model IDs tongyi-embedding-vision-flash, tongyi-embedding-vision-plus, and multimodal-embedding-v1, respectively (Alibaba Cloud, 2026).

Datasets. We evaluate models on three benchmarks. CulTi2.0 is our main benchmark for Cultural Boolean Retrieval, where queries are constructed from typed cultural entities and attributes. We further evaluate on two general-domain negation benchmarks. NegCOCO (Park et al., 2025) evaluates negation understanding in image-text matching: given a negation-inclusive prompt, the model selects the matching image from two candidates. The Existence split of VALSE (Parcalabescu et al., 2022) evaluates object presence and absence: given an image and two textual candidates, one positive and one negated, the model selects the caption that best matches the image.

5.2 Comparison with Existing Models

We evaluate CLAIR against mainstream multimodal embeddings on our CulTi2.0 and two general-domain negation benchmarks: VALSE (Parcalabescu et al., 2022) and NegCOCO (Park et al., 2025). Overall, CLAIR consistently achieves state-of-the-art performance across all datasets. As shown in Fig. 4, CLAIR boosts Qwen3-VL-Emb (Li et al., 2026)'s accuracy on VALSE from 65.35% to 81.98%. On CulTi2.0, it improves Qwen3-VL-Emb (Li et al., 2026)'s image-to-text RSum from 19.24% to 31.05% on the *Entity AND Attribute* subset. Detailed retrieval results are reported in Table 1.

Consistent Enhancements in Logical Understanding. CLAIR systematically addresses baseline logical shortcomings. On CulTi2.0, Table 1 shows that CLAIR increases text-to-image RSum from 10.59% to 11.81% (*Entity AND Entity*) and 11.66% to 14.04% (*Entity AND Attribute*), suggesting explicit operator-aware scoring is vital for binding fine-grained attributes. These gains also generalize beyond CulTi2.0: as shown in Fig. 4, Tongyi-Plus (Alibaba Cloud, 2026) improves by 11.29% on VALSE, while Qwen3-VL-Emb (Li et al., 2026) improves from 65.23% to 74.09% on NegCOCO.

The asymmetric gains reflect the different roles of positive and negative evidence. AND queries provide positive evidence for both concepts, whereas NOT queries require suppressing evidence for the excluded concept. Holistic embeddings may therefore remain partially aligned with AND queries but mis-rank NOT queries.

Robustness Across Model Families. CLAIR's

Model	Entity AND Entity				Entity AND Attribute				Entity NOT Entity				Entity NOT Attribute			
	R@1	R@5	R@10	RSum	R@1	R@5	R@10	RSum	R@1	R@5	R@10	RSum	R@1	R@5	R@10	RSum
Text-to-Image Retrieval																
ALBEF (Li et al., 2021)	0.00	0.06	0.10	0.17	0.02	0.11	0.12	0.25	0.61	2.27	4.05	6.93	0.05	0.50	1.32	1.87
BLIP (Li et al., 2022)	0.07	0.25	0.41	0.73	0.00	0.00	0.15	0.15	0.78	3.31	6.32	10.41	0.12	0.60	2.27	2.99
CN-CLIP (Yang et al., 2022)	1.13	3.74	5.71	10.59	1.18	4.05	6.25	11.49	6.99	22.41	31.88	61.28	5.74	15.74	23.23	44.71
MM-Emb-v1 (Alibaba Cloud, 2026)	0.87	3.01	4.88	8.76	0.92	3.30	5.22	9.44	6.56	20.10	30.36	57.02	3.23	13.28	20.93	37.44
Qwen3-VL-Emb (Li et al., 2026)	0.89	2.95	4.58	8.42	1.46	4.14	6.06	11.66	8.71	25.46	35.29	69.47	3.47	14.59	22.73	40.79
SigLIP2 (Tschannen et al., 2025)	0.86	2.75	4.36	7.97	0.92	2.76	5.14	8.82	6.13	20.83	29.93	56.88	4.19	14.35	20.10	38.64
Tongyi-Flash (Alibaba Cloud, 2026)	0.33	1.22	1.99	3.54	0.54	1.77	2.76	5.07	5.70	19.79	28.34	53.83	2.39	10.29	15.79	28.47
Tongyi-Plus (Alibaba Cloud, 2026)	0.32	1.33	2.32	3.97	0.69	1.61	3.53	5.83	5.97	18.40	27.56	51.93	3.83	13.04	17.70	34.57
CLAIR(OURS)	**1.21**	**4.17**	**6.44**	**11.81**	**1.77**	**4.76**	**7.52**	**14.04**	**11.49**	**30.55**	**38.61**	**80.64**	**7.78**	**18.66**	**24.04**	**50.48**
Image-to-Text Retrieval																
ALBEF (Li et al., 2021)	0.03	0.08	0.18	0.29	0.00	0.15	0.59	0.74	0.58	2.91	4.05	6.93	0.67	2.05	4.06	6.78
BLIP (Li et al., 2022)	0.08	0.10	0.28	0.46	0.15	0.67	1.56	2.38	1.10	4.27	7.90	13.27	0.67	4.68	8.43	13.78
CN-CLIP (Yang et al., 2022)	1.65	4.19	5.92	11.77	0.98	3.51	5.50	9.99	4.29	16.68	25.85	46.82	4.75	10.63	15.21	30.60
MM-Emb-v1 (Alibaba Cloud, 2026)	1.11	3.70	5.63	10.44	1.34	4.53	8.32	14.19	6.24	20.63	28.35	55.22	3.79	13.25	17.31	34.35
Qwen3-VL-Emb (Li et al., 2026)	1.71	5.51	8.43	15.65	1.63	6.76	10.85	19.24	5.57	20.13	**31.46**	57.17	5.75	14.99	**21.28**	42.02
SigLIP2 (Tschannen et al., 2025)	0.96	3.26	5.01	9.23	1.26	3.94	6.69	11.89	5.49	17.85	30.18	53.52	5.71	14.76	20.34	40.81
Tongyi-Flash (Alibaba Cloud, 2026)	0.65	2.02	3.44	6.11	1.34	4.75	7.95	14.04	3.87	12.31	20.35	36.53	2.50	9.77	15.74	28.01
Tongyi-Plus (Alibaba Cloud, 2026)	1.03	3.31	5.48	9.82	2.01	6.32	10.03	18.36	5.37	15.34	23.64	44.36	2.94	12.53	19.58	35.05
CLAIR(OURS)	**2.09**	**7.42**	**10.91**	**20.42**	**3.64**	**10.85**	**16.57**	**31.05**	**12.41**	**22.36**	28.25	**63.02**	**8.03**	**15.70**	20.61	**44.34**

Table 1: Retrieval results on CulTi2.0. We report Recall@K (%) and RSum, computed from Recall@1, Recall@5, and Recall@10. The **best** and second-best results within each retrieval task (Text-to-Image and Image-to-Text) are highlighted in bold and underline, respectively. OURS is implemented with relation-aware scoring and negative projection based on Qwen3-VL-Emb.

Dataset	Model	Correct		Incorrect	
		Orig.	Ours	Orig.	Ours
VALSE	MM-Emb-v1 (Alibaba Cloud, 2026)	62.96	78.44	37.04	21.56
	Qwen3-VL-Emb (Li et al., 2026)	64.66	80.11	35.34	19.89
	Tongyi-Flash (Alibaba Cloud, 2026)	67.74	84.15	32.26	15.85
	Tongyi-Plus (Alibaba Cloud, 2026)	68.55	85.09	31.45	14.91
NegCOCO	MM-Emb-v1 (Alibaba Cloud, 2026)	57.69	63.36	42.31	36.64
	Qwen3-VL-Emb (Li et al., 2026)	63.09	71.06	36.91	28.94
	Tongyi-Flash (Alibaba Cloud, 2026)	60.09	66.09	39.91	33.91
	Tongyi-Plus (Alibaba Cloud, 2026)	58.54	64.51	41.46	35.49

Table 2: Preference redistribution analysis on general-domain negation benchmarks.

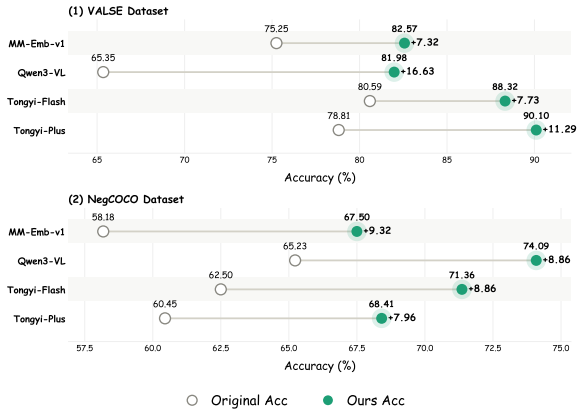


Figure 4: Performance comparison on VALSE and NegCOCO. Each arrow shows the change in accuracy from the original embedding model to CLAIR.

gains are highly architecture-agnostic. Table 1 shows that strong baseline semantic alignment does not guarantee robust Boolean retrieval ability: CN-CLIP (Yang et al., 2022) excels in text-to-image with an RSum of 10.59% on *Entity AND Entity*, while Qwen3-VL-Emb (Li et al., 2026) dominates image-to-text, achieving 19.24% on *Entity AND Attribute*. Despite these baseline variations, CLAIR consistently boosts logical reasoning performance. As shown in Fig. 4, these substantial improvements hold across both open-source models and closed-source commercial APIs. Most notably, on complex logical setups such as *Entity NOT Attribute*, CLAIR drives impressive RSum improvements of up to +4.91% on Tongyi-Flash (Alibaba Cloud, 2026) and +3.21% on Tongyi-Plus (Alibaba Cloud, 2026) in image-to-text retrieval. This confirms CLAIR as a highly effective plug-and-play correction strategy without requiring model retraining.

5.3 Ablation Study

We conduct comprehensive ablation and preference shift analyses to validate CLAIR. Overall, results confirm that all core components act synergistically to boost retrieval accuracy and fundamentally correct the model's logical preferences.

Complementary Contributions of Logic-Aware Components. Table 3 reports our step-wise ablations. Logical operator scoring yields immediate baseline improvements, proving the efficacy of continuous Boolean computation. Notably, it drives accuracy on VALSE from 75.00% to 80.25%, indicating that models benefit heavily from explicit operator grounding. Modifier relation scoring provides further steady gains (e.g., increasing CulTi2.0 RSum from 28.36% to 29.47%) by strictly binding entities to attributes rather than treating them as fragmented cues. Furthermore, negative evidence suppression drives a crucial performance leap on negation-focused benchmarks, boosting

NegCOCO accuracy from 66.08% to 68.72% and VALSE to 83.96% by actively punishing logically invalid concepts. Ultimately, the full CLAIR framework achieves the best performance across all datasets (e.g., 31.89% on CulTi2.0 and 85.74% on VALSE), confirming that all proposed components are highly complementary.

Enhancing Logical Distinguishability. Beyond correcting isolated predictions, CLAIR fundamentally widens the gap between logically valid and invalid candidates. To quantify this, we apply a temperature-scaled softmax to convert narrow, model-dependent cosine similarities into probabilistic preferences, strictly preserving original rankings (Table 2). Consequently, the average preference for valid candidates surges significantly: from 65.98% to 81.95% on VALSE, and 59.85% to 66.26% on NegCOCO. This confirms that effective Boolean retrieval necessitates explicitly suppressing logically contradictory evidence to make correct candidates unambiguously distinguishable.

Method	CulTi2.0	VALSE	NegCOCO
Baseline	26.78	75.00	61.59
+ Logical Operator Scoring	28.36	80.25	65.66
+ Modifier Relation Scoring	29.47	80.92	66.08
+ Negative Evidence Suppression	31.02	83.96	68.72
CLAIR	**31.89**	**85.74**	**70.34**

Table 3: Ablation study of CLAIR. We report weighted average RSum (%) on CulTi2.0 and accuracy (%) on VALSE and NegCOCO.

5.4 Case Study

We further construct multi-term Boolean queries to evaluate whether CLAIR can generalize beyond simple pairwise retrieval. As shown in Fig. 5, both examples involve ternary logical composition. Case 1 requires simultaneously satisfying three cultural entities, while Case 2 requires preserving one entity while excluding two semantically related entities. In both cases, CLAIR substantially improves retrieval quality. In Case 1, the target rank improves from 4,153 to 134, with the retrieval score increasing from 0.2554 to 0.5783. In Case 2, the target image is moved from rank 5,723 to rank 2, with a score increase from 0.0745 to 0.2220. These examples suggest that relation-aware scoring remains effective under more compositional logical retrieval settings.

5.5 Findings on Existing Embedding Models

Through our extensive evaluation on CulTi2.0 and general-domain datasets, we identify two critical limitations in current multimodal embedding spaces regarding logical comprehension.

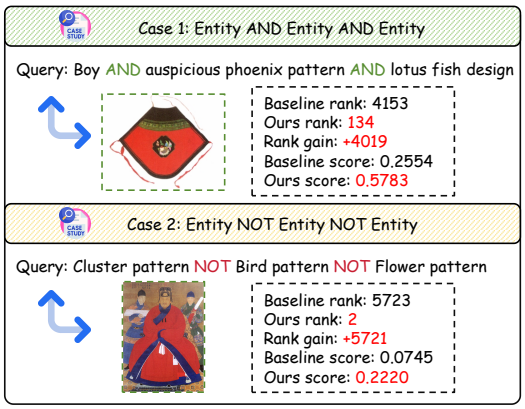


Figure 5: Case studies of CLAIR on ternary Boolean retrieval queries. The baseline is Qwen3-VL-Emb.

Observation 1: Vulnerability to Cultural Boolean Constraints. All evaluated baseline models achieve strictly low absolute recall on CulTi2.0. For text-to-image retrieval, the best baseline RSum is only 10.59% on *Entity AND Entity*, 11.66% on *Entity AND Attribute*, and 44.71% on *Entity NOT Attribute*. We attribute this poor performance to a compounded challenge: general-purpose embeddings lack deep cultural knowledge, and this domain gap is severely exacerbated by their inherent inability to handle fine-grained Boolean composition.

Observation 2: Systemic Flaws in General Semantic Exclusion. This logical failure extends well beyond the cultural domain. Across public benchmarks, baseline embeddings frequently confuse semantically related positive and negated descriptions. For instance, Qwen3-VL-Emb (Li et al., 2026) achieves only 65.35% accuracy on VALSE and 65.23% on NegCOCO, while MM-Emb-v1 (Alibaba Cloud, 2026) scores merely 58.18% on NegCOCO. This indicates that current methods natively treat mentioned concepts as positive matching evidence, struggling fundamentally when queries require explicit semantic exclusion.

6 Conclusion

In this paper, we introduced Multimodal Boolean Logic Culture Retrieval, a novel task that challenges models to satisfy explicit Boolean constraints rather than relying on mere semantic overlap. To systematically study this problem, we constructed CulTi2.0, a dedicated benchmark for typed cultural Boolean retrieval. Furthermore, we proposed CLAIR, a training-free, logic-aware framework that enforces hard logical constraints directly

within frozen multimodal embedding spaces. Extensive experiments across CulTi2.0 and general-domain negation datasets demonstrate that while current embeddings fundamentally struggle with logical validity, CLAIR serves as an effective plug-and-play solution that consistently boosts performance across diverse model families.

Limitations

Because many widely used embedding models are closed-source, we focus on a training-free correction framework that operates over frozen embeddings. Future work could develop Boolean-aware embedding models that internalize logical structure through logic-violating hard negatives, entity-attribute binding supervision, and logic-consistency objectives.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under No. 62276258 and the Jiangsu Science and Technology Programme BK20251812, and the Zhejiang Province Leading Earth Goose Program 2026LDC01028 (XC).

References

Alibaba Cloud. Multimodal embedding API [online]. 2026.

Jaeseok Byun, Young Kyun Jang, Seokhyeon Jeong, Donghyun Kim, and Taesup Moon. 2025. Ma-cir: A multimodal arithmetic benchmark for composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21342–21352.

Valentina Anita Carriero, Aldo Gangemi, Maria Letizia Mancinelli, Ludovica Marinucci, Andrea Giovanni Nuzzolese, Valentina Presutti, and Chiara Veninata. 2019. Arco: The italian cultural heritage knowledge graph. In *International semantic web conference*, pages 36–52. Springer.

Tao Fan and Hao Wang. 2022. Research of chinese intangible cultural heritage knowledge graph construction and attribute value extraction with graph attention network. *Information Processing & Management*, 59(1):102753.

Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, 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. In *ACL Findings*, pages 23627–23641.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Prachi Jha, Sumit Bhatia, and Srikanta Bedathur. 2026. Being positive about negative queries: Exclusion aware multimodal retrieval using disentangled representations. In *WACV*, pages 7832–7841.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781.

George Klir and Bo Yuan. 1995. *Fuzzy sets and fuzzy logic*, volume 4. Prentice hall New Jersey.

Taegyeong Lee, Jiwon Park, Seunghyun Hwang, and JooYoung Jang. 2026. Deo: Training-free direct embedding optimization for negation-aware retrieval. *arXiv preprint arXiv:2603.09185*.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and 1 others. 2025. Ravenea: A benchmark for multimodal retrieval-augmented visual culture understanding. *arXiv preprint arXiv:2505.14462*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, and 1 others. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *ACL*, pages 8253–8280.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. 2025. Know "no" better: A data-driven approach for enhancing negation awareness in clip. In *ICCV*, pages 2825–2835.

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 *ICML*, pages 8748–8763.

Hongyu Ren and Jure Leskovec. 2020. Beta embeddings for multi-hop logical reasoning in knowledge graphs. *NeurIPS*, 33:19716–19726.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Chih-Fong Tsai. 2007. A review of image retrieval methods for digital cultural heritage resources. *Online Information Review*, 31(2):185–198.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Coen van den Elsen, Francien Barkhof, Thijmen Nijdam, Simon Lupart, and Mohammad Aliannejadi. 2025. Reproducing nevir: Negation in neural information retrieval. In *ACM SIGIR*, pages 3346–3356.

Vladi Vexler, Ofer Idan, Gil Lederman, and Dima Sivov. 2026. Nsfl: A post-training neuro-symbolic fuzzy logic framework for boolean operators in neural embeddings. *arXiv preprint arXiv:2604.10604*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *ICLR*.

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*.

Junyi Yuan, Jian Zhang, Fangyu Wu, Huanda Lu, Dongming Lu, and Qiufeng Wang. 2025. Towards cross-modal retrieval in chinese cultural heritage documents: Dataset and solution. In *ICDAR*, pages 570–586. Springer.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986.

Jian Zhang, Junyi Guo, Junyi Yuan, Huanda Lu, Yanlin Zhou, Fangyu Wu, Qiufeng Wang, and Dongming Lu. 2025. Llm-driven completeness and consistency evaluation for cultural heritage data augmentation in cross-modal retrieval. In *EMNLP*, pages 19418–19428.

Zhanqiu Zhang, Jie Wang, Jiajun Chen, Shuiwang Ji, and Feng Wu. 2021. Cone: Cone embeddings for multi-hop reasoning over knowledge graphs. *NeurIPS*, 34:19172–19183.

Zongmeng Zhang, Jinhua Zhu, Wengang Zhou, Xiang Qi, Peng Zhang, and Houqiang Li. 2024. Boolquestions: Does dense retrieval understand boolean logic in language? In *EMNLP Findings*, pages 2767–2779.