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

Yuchen Huang[1], Zhiyuan Fan[1], Zhitao He[1], Sandeep Polisetty[2], Wenyan Li[3]
Yi R. (May) Fung[1]
[1]Hong Kong University of Science and Technology
[2]University of Massachusetts Amherst, [3]University of Copenhagen
{yhuanggn, yrfung}@cse.ust.hk

Abstract

Pretrained vision-language models (VLMs) such as CLIP excel in general multimodal comprehension but often struggle to capture nuanced, context-dependent visual cues. This makes it difficult to distinguish between similar-looking concepts with potentially different cultural meanings. Such deficiencies are mainly due to a limited amount of high-quality cultural data, contextual information, and the lack of negative examples that highlight subtle differences. To mitigate this, we design a data curation pipeline leveraging open-sourced VLMs and text-to-image models to construct **CulTwin**, a synthetic cultural dataset. This dataset consists of paired concept-caption-image triplets, where concepts visually resemble each other but are culturally different. Then, we fine-tune CLIP on CulTwin to develop **CultureCLIP**, which aligns cultural concepts with contextually enhanced captions and synthetic images through tailored contrastive learning. Experiments on culture-specific benchmarks show that CultureCLIP outperforms the base CLIP, achieving up to a notable 5.49% improvement in fine-grained concept recognition on certain tasks while preserving CLIP's original generalization ability, validating the effectiveness of our data synthesis and VLM backbone training paradigm in capturing subtle cultural distinctions.[1]

1 Introduction

Recent advancements in vision-language reasoning (Zhang et al., 2024b; Lu et al., 2024) have revolutionized multimodal understanding by efficiently integrating visual and linguistic semantics within a shared feature space (Jia et al., 2021; Radford et al., 2021). By leveraging large-scale image-text corpora, models such as Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) exhibit remarkable generalization capabilities across diverse downstream tasks, including visual question answering (VQA)(Shen et al., 2021; Song et al., 2022), cross-modal retrieval (Koukounas et al., 2024; Baldrati et al., 2023), and zero-shot image classification (Zhou et al., 2022; Saha et al., 2024). CLIP is built on a contrastive learning objective, where two separate encoders are trained to bring matching image-text pairs closer in the feature space while pushing apart non-matching pairs within the same batch. Due to the concise nature of the text in its training data, CLIP is effective at coarse-grained semantic alignment, particularly in identifying the general type of the main object (Radford et al., 2021; Zhang et al., 2024a). However, it often struggles with fine-grained alignment, especially when contextual visual details, such as specific accessories, stylistic cues, or symbolic elements that convey meaning only within particular cultural contexts, are required to distinguish between visually similar but culturally distinct concepts.

For example, CLIP might correctly identify both *Yuelao* (the Chinese god of love and marriage) and *Taishang Laojun* (a Daoist patriarch) as elderly Chinese deities, but it often

[1]Our code is publicly available at `https://github.com/lukahhcm/CultureCLIP`.

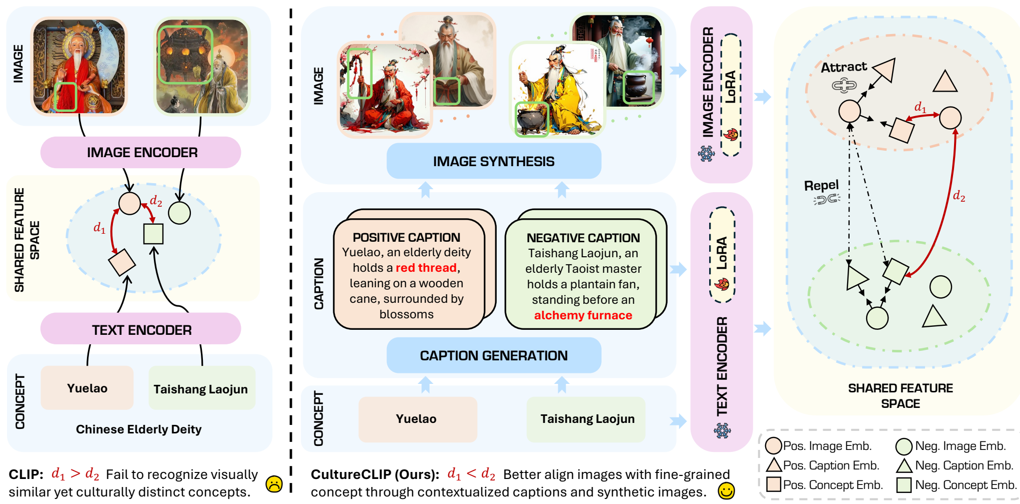

Figure 1: CLIP vs. CultureCLIP. **Left:** The original CLIP model fails to capture fine-grained contextual visual cues (highlighted by the green box), leading to mismatches with cultural concepts. Although both the image and concept projections fall within the region outlined by the purple dashed box—i.e., the semantic space of Chinese elderly deities—the distance between the image of Yuelao (pink circle) and its correct concept (pink square) is greater than that to an incorrect one such as Taishang Laojun (blue square), i.e., $d_1 > d_2$. **Right:** CultureCLIP improves fine-grained cultural understanding by jointly aligning concepts with contextualized captions and their corresponding synthetic images, while repelling unrelated concepts and captions in the embedding space (see details in Section 4.2).

struggles to tell them apart because of its insensitivity to capturing subtle yet crucial visual details, like the *red thread* for *Yuelao*, which stands for love, or the *alchemy furnace* for *Taishang Laojun*, which stands for immortality (both shown in green in Figure 1). These culturally specific visual cues, however, are exactly what people in particular cultural groups use to distinguish fine-grained concepts. This raises an intriguing question: **How can we teach CLIP to capture such details so that it can differentiate between cultural concepts that share visual similarities?**

A natural approach is to curate a large-scale dataset comprising visually similar cultural concept pairs (i.e., original concepts alongside their hard negatives) accompanied by image-text pairs enriched with contextual cultural knowledge and illustrating subtle visual distinctions. However, collecting such data presents three major challenges: **First, high-quality cultural image-text pairs are scarce and costly to annotate.** Existing manually curated cultural datasets are typically limited in scale, often containing only a few thousand samples (Nikandrou et al., 2024; Bhatia et al., 2024; Romero et al., 2024), due to the labor-intensive nature of data collection and annotation, as well as the need for domain-specific expertise. While web-scraped data might serve as an alternative (Xu et al., 2023), it frequently introduces substantial noise: images may be ethically inappropriate, copyright-protected, low-resolution, or culturally misleading, and the corresponding text may be mismatched or entirely absent. **Second, CLIP's original training strategy inherently favors concise captions.** Specifically, it imposes a 77-token limit, with the majority of alignment achieved within the first 20 tokens (Zhang et al., 2024a). This design makes it particularly challenging to incorporate lengthy, information-rich cultural context directly into the text for training, as doing so may disrupt the original image-text alignment learned by the model. **Third, fine-grained hard negatives specifically tailored for visually similar concepts are lacking.** In the original CLIP framework, negative samples are randomly drawn within each batch, which significantly reduces the likelihood of including conceptually similar but visually distinct examples. Although recent works (Yuksekgonul et al., 2022; Patel et al., 2024) introduce harder negatives by making minor modifications to captions or images, they still lack structured negative samples that highlight both coarse-grained similarities and fine-grained differences, which are crucial for enabling culturally grounded visual distinctions.

Motivated by these challenges, we introduce **CulTwin**, a synthetic dataset of paired concept-caption-image triplets, which we call *Twin Cards*. In each card, two similar concepts are paired together, and their captions are enriched with cultural background knowledge using a vision-language model. The corresponding images are then generated from these captions by a text-to-image model. Building on CulTwin, we propose **CultureCLIP**, a contrastive learning framework that jointly aligns concepts, captions, and images in a shared embedding space. As illustrated in Figure 1 (right), our training objective works like a magnetic field: it attracts each concept toward its corresponding caption and image, while repelling it from the captions and images of its culturally contrasting counterpart. Experiments on culture-specific benchmarks show that CultureCLIP significantly outperforms the original CLIP model, achieving over 5% improvement in fine-grained concept recognition on specific tasks. These results highlight the effectiveness of our synthetic dataset and training methodology in capturing subtle cultural nuances.

2 Related Work

Advancements in Vision-Language Models Recent studies on multimodal reasoning (Peng et al., 2025; Zhao et al., 2025; Zhang et al., 2025; He et al., 2025) have significantly advanced vision-language models and broadened their capabilities across various downstream tasks and domains. For instance, Med-Flamingo (Moor et al., 2023) unlocks medical VQA abilities through continued pre-training on paired and interleaved medical image-text data, while ChemVLM (Li et al., 2024), trained on bilingual image-text data, enhances the joint understanding of textual and visual chemical information. In the cultural domain, CultureVLM (Liu et al., 2025) improves cultural understanding by fine-tuning on a large-scale multimodal benchmark, CultureVerse. Despite these advancements, existing vision-language models still struggle to capture fine-grained visual cues and often misclassify visually similar but culturally distinct concepts. We propose CultureCLIP, which aligns cultural concepts with enriched captions and synthetic images through contrastive learning, improving cultural differentiation while preserving generalization capabilities.

Data for Vision-Language Pre-training Cross-modal mutual information maximization relies on large-scale, diverse training data that captures real-world concepts and relationships. With more than 5 billion internet-derived image-caption pairs, the LAION dataset (Schuhmann et al., 2022) serves as a critical training resource. MetaCLIP (Xu et al., 2023) formalizes CLIP's implicit data selection via explicit metadata balancing, creating 400M CommonCrawl pairs. SynthCLIP (Hammoud et al., 2024) reduces reliance on web-scraped data by generating over 30 million synthetic pairs. LaCLIP (Fan et al., 2023) enhances text augmentation through in-context language model rewriting. For human-centric AI, fine-grained cultural understanding is essential, yet culturally relevant multimodal data is scarce. While Nayak et al. (2024); Liu et al. (2025) introduced cultural datasets, the diversity of images hinders vision-language models from learning cultural distinctions. In this work, we construct CulTwin, a synthetic dataset comprising concept-caption-image triplets enriched with cultural contextual knowledge.

Contrastive Pre-training Contrastive learning has become a strong method for multimodal representation learning, with CLIP (Radford et al., 2021) demonstrating scalability and zero-shot transfer potential. More efficient contrastive pre-training methods have been proposed for finer-grained multimodal representations learning (Zhang et al., 2024c; Patel et al., 2024). BLIP-2 (Li et al., 2023) introduces a lightweight Querying Transformer for cost-efficient pre-training. NegCLIP (Yuksekgonul et al., 2022) generates hard negative captions through semantic perturbations. TripletCLIP (Patel et al., 2024) uses hard negative pairs with a triplet contrastive loss. Existing contrastive pre-training methods focus on caption-image pairs and their negative samples, but fail to capture culturally relevant information due to its multidimensional nature. To address this, CultureCLIP enhances cultural understanding

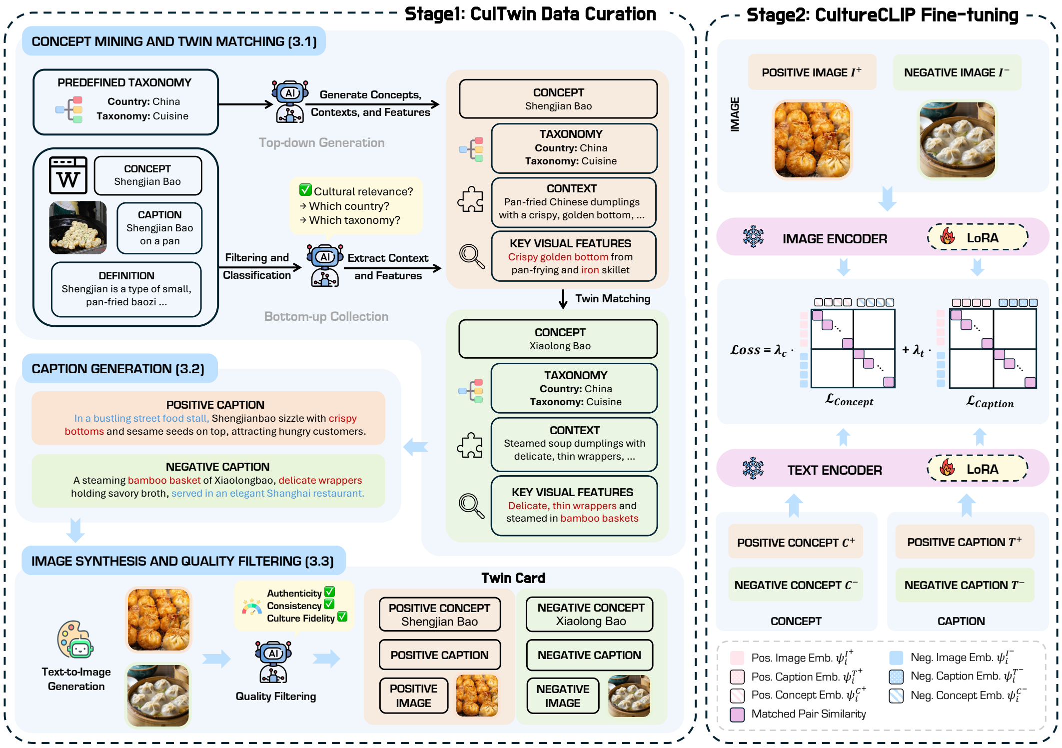

Figure 2: **Left:** Data curation pipeline in CulTwin. **Right:** Architecture of CultureCLIP.

by aligning cultural concepts with contextualized captions and synthetic images, while separating unrelated concepts in the embedding space.

3 CulTwin: A Three-stage Cultural Data Curation Pipeline

High-quality and diverse data has been shown to be essential for training models like CLIP (Nguyen et al., 2023; Fang et al., 2022). In this section, we present a three-stage data curation pipeline for constructing **CulTwin**, a synthetic cultural dataset composed of *Twin Cards*—pairs of concept-caption-image triplets that are visually similar but culturally distinct. The pipeline begins by collecting culturally grounded concepts, their background knowledge, and visual features, and performing twin matching to identify negative samples that are visually similar but culturally different (Section 3.1). Next, diverse captions are generated by leveraging cultural context and key visual features through a large language model (LLM) (Section 3.2). Finally, images are synthesized from the captions and evaluated using a Vision-Language Model (VLM), which scores each image based on authenticity, consistency, and cultural fidelity to guide data quality filtering (Section 3.3). The resulting concept-caption-image triplets are then organized into *Twin Cards* for later fine-grained contrastive training. Figure 2 (Left) provides an overview of the full CulTwin data curation pipeline.

3.1 Concept Mining and Twin Matching

We begin with a manually predefined taxonomy covering 229 countries and 8 cultural categories, including *Cuisine, Clothing, Animals & Plants, Art, Architecture, Daily Life, Symbols, and Festivals*. This taxonomy is designed to capture a broad and representative set of cultural elements (definitions of each category are provided in Appendix A). We then collect culturally meaningful concept candidates through both bottom-up collection and top-down generation.

Table 1: Filtering Outcomes and Image Quality Scores Across Diverse Cultural Taxonomies

Taxonomy	Image Quality Scores After Filtering						Filtering Outcomes	
	MLLM-as-a-Judge			Human Evaluation			Pass %	Retained
	Auth ↑	Cons ↑	Fid ↑	Auth ↑	Cons ↑	Fid ↑		
Cuisine	4.403 ± 0.528	3.548 ± 0.627	3.750 ± 0.719	4.517 ± 0.866	3.533 ± 0.957	3.550 ± 1.023	77.46	24824/32,046
Clothing	4.469 ± 0.397	3.862 ± 0.422	4.002 ± 0.152	4.550 ± 0.956	3.483 ± 1.218	3.433 ± 1.202	75.52	7250/9,600
Animal & Plants	4.350 ± 0.415	3.979 ± 0.528	4.284 ± 0.366	4.183 ±1.162	3.783 ± 1.185	3.800 ± 1.137	75.63	10966/14,500
Art	4.390 ± 0.444	4.032 ± 0.415	4.093 ± 0.253	4.433 ± 0.761	2.950 ± 1.189	2.917 ± 1.144	72.67	15806/21,750
Architecture	4.222 ± 0.349	3.897 ± 0.402	3.997 ± 0.335	4.183 ± 1.025	3.117 ± 1.112	2.917 ± 1.100	61.32	6377/10,400
Daily Life	4.239 ± 0.436	3.882 ± 0.489	4.152 ± 0.308	4.483 ± 0.904	4.050 ± 1.132	4.100 ± 1.179	72.80	6188/8,500
Symbol	3.802 ± 0.371	3.830 ± 0.291	3.977 ± 0.109	4.467 ± 0.718	3.950 ± 1.310	3.833 ± 1.227	85.12	681/800
Festival	3.971 ± 0.326	3.921 ± 0.332	3.994 ± 0.061	4.167 ± 0.986	3.533 ± 1.087	3.533 ± 1.008	72.12	1731/2,400

Scores are averaged over three dimensions—**Auth** (image authenticity), **Cons** (concept consistency), and **Fid** (cultural fidelity)—rated from 1 to 5. **Pass %** indicates the proportion of images passing quality thresholds; **Retained** shows the number of remaining samples per category.

Bottom-up Concept Collection In this approach, we first collect candidate concepts and their background information (definitions, images, captions) from Wikipedia. We then use Qwen2.5-VL (Bai et al., 2025) to assess each concept's cultural relevance, discarding any that are not strongly related to the predefined categories. To ensure strict filtering, we designed prompts (Appendix F) that favor rejection over uncertain inclusion, prioritizing data quality over quantity. For the retained concepts, we assign metadata (country and cultural category) and extract key contextual and visual features.

Top-down Concept Generation In this approach, we further expand the concept pool using the predefined taxonomy. For each cultural category and country, Qwen2.5-VL is employed to generate culturally grounded concepts, along with associated context and visual features.

After obtaining these filtered concepts from both approaches, we perform *twin matching* to identify culturally distinct but visually similar hard negatives for each concept, using Qwen2.5-VL conditioned on its context and key visual features. These final concept pairs, together with their metadata and contextual information, form the foundation for generating detailed captions and images in subsequent stages.

3.2 Diverse Caption Generation

In the second stage, we generate diverse, culturally rich captions for each concept in the paired sets. These captions are designed to highlight both key visual elements and cultural nuances, preserving subtle distinctions before image synthesis. We leverage Qwen2.5-VL to incorporate cultural context and salient visual features, producing contextualized descriptions. To mitigate the risk of limited diversity and potential overfitting in synthetic data (Hammoud et al., 2024), we guide the model to vary aspects such as artistic style, scene setting, and compositional details, thereby enriching the visual representation of each concept.

3.3 Image Synthesis and Quality Filtering

After caption generation, we synthesize images using Stable Diffusion 3.5 (Rombach et al., 2022), followed by quality filtering with MLLM-as-a-Judge (Chen et al., 2024), implemented using Qwen-VL-2.5. Each synthesized image is evaluated across three key dimensions on a scale from 1 to 5:

1) Authenticity: Evaluates the physical realism and adherence to common human understanding. **2) Consistency:** Assesses alignment between the image and its caption, ensuring accurate concept representation. **3) Cultural Fidelity:** Examines the preservation and correctness of cultural features specific to the concept.

Images receiving a score of 1 in any dimension or an average score below 3 are discarded. To validate the automated filtering, we also perform a human evaluation on a sampled subset,

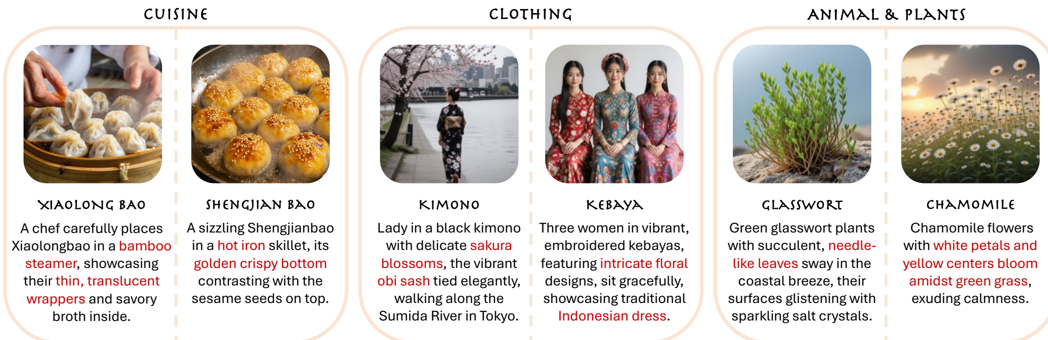

Figure 3: *Twin Card* examples from the *Cuisine*, *Clothing*, and *Animals & Plants* categories. Each Twin Card contrasts two culturally distinct yet visually similar concepts, with each side comprising a concept-caption-image triplet (see Section 3.3). Additional examples are provided in Appendix B.

where three PhD-level experts independently score the images using the same criteria. Summary statistics of the automated and human evaluations are shown in Table 1.

Categories like *Cuisine* and *Clothing* exhibit high **Authenticity** scores, reflecting their tangible and universally recognized nature, making them easier to evaluate accurately by both models and human judges. In contrast, more abstract categories such as *Festival* tend to have lower **Authenticity** and **Cultural Fidelity** scores, particularly in human evaluations, due to their complex and diverse cultural elements. Furthermore, *Art* and *Architecture* show a noticeable gap between automated and human evaluations, especially in **Consistency** and **Cultural Fidelity**. These categories involve nuanced cultural and conceptual details that automated models struggle to capture, highlighting the importance of human judgment for these intricate domains.

Each final *Twin Card* is constructed by assembling two triplets—each consisting of a concept, its caption (from Section 3.2), and the corresponding synthesized image—to explicitly highlight cultural contrast while maintaining visual similarity. Example Twin Cards are illustrated in Figure 3, and additional details of CulTwin can be found in Appendix B.

4 CultureCLIP: Fine-Grained Cultural Alignment

4.1 Preliminary

CLIP learns joint image-text representations via an image encoder $\mathcal{F} : \mathcal{I} \to \mathbb{R}^d$ and a text encoder $\mathcal{G} : \mathcal{T} \to \mathbb{R}^d$, projecting inputs into a shared embedding space $\mathcal{V}$ of dimension d. Given a batch of N image-text pairs $\{(I_i, T_i)\}_{i=1}^N$, representations are computed as $\psi_i^I = \mathcal{F}(I_i)$ and $\psi_i^T = \mathcal{G}(T_i)$. CLIP constructs a similarity matrix $S \in \mathbb{R}^{N\times N}$, where each entry $S_{i,j} = \mathrm{sim}(\psi_i^I, \psi_j^T)$ denotes the cosine similarity between image I_i and text T_j. The contrastive loss encourages each aligned pair (on the diagonal of S) to have a higher similarity than all mismatched pairs in the same row or column. Formally:

$$\mathcal{L}_{\text{CLIP}}(I, T) = \mathcal{L}_{\text{I2T}}(I, T) + \mathcal{L}_{\text{T2I}}(I, T), \tag{1}$$

$$\mathcal{L}_{\text{I2T}}(I, T) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(\psi_i^I, \psi_i^T)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(\psi_i^I, \psi_k^T)/\tau)}, \tag{2}$$

$$\mathcal{L}_{\text{T2I}}(I, T) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathrm{sim}(\psi_i^I, \psi_i^T)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(\psi_k^I, \psi_i^T)/\tau)}. \tag{3}$$

Here, τ is a temperature parameter that controls the sharpness of the similarity distribution.

NegCLIP (Yuksekgonul et al., 2022) builds on this framework by introducing hard negative captions T^-, derived through semantic perturbations of the original texts T^+. These hard negatives are added alongside the standard in-batch negatives, forming an extended candidate set $\tilde{T} = \{T^+\} \cup \{T^-\}$ and resulting in a similarity matrix $\tilde{S} \in \mathbb{R}^{N \times 2N}$. The contrastive objective is thus extended beyond standard in-batch negatives to also include explicitly constructed hard negatives:

$$\mathcal{L}_{\text{NegCLIP}}(I, T^+, T^-) = \mathcal{L}_{\text{I2T\_neg}}(I, T^+, T^-) + \mathcal{L}_{\text{T2I}}(I, T^+), \tag{4}$$

$$\mathcal{L}_{\text{I2T\_neg}}(I, T^+, T^-) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\psi_i^I, \psi_i^{T^+})/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\psi_i^I, \psi_k^{T^+})/\tau) + \sum_{m=1}^{N} \exp(\text{sim}(\psi_i^I, \psi_m^{T^-})/\tau)}. \tag{5}$$

TripletCLIP (Patel et al., 2024) further incorporates hard negative images I^-. The training objective encourages I^+ to be closer to T^+ than to T^-, and symmetrically, I^- to align more with T^- than with T^+. This is achieved by summing two NegCLIP-style losses:

$$\mathcal{L}_{\text{TripletCLIP}}(I^+, I^-, T^+, T^-) = \mathcal{L}_{\text{NegCLIP}}(I^+, T^+, T^-) + \mathcal{L}_{\text{NegCLIP}}(I^-, T^-, T^+). \tag{6}$$

While prior work enhances image-text alignment using modality-specific hard negatives, it tends to focus on coarse-grained semantic differences. We extend this framework by introducing abstract concepts as anchors to better connect detailed captions and images. Through a refined training objective (Section 4.2), our model intends to capture fine-grained cultural semantics with greater precision.

4.2 CultureCLIP

To better capture fine-grained cultural semantics, we build on the *Twin Cards* introduced in **Cultwin**, where each card contains two triplets—(C^+, T^+, I^+) and (C^-, T^-, I^-)—representing visually similar but culturally distinct concepts. These pairs serve as hard negatives for each other. Our goal is to align each concept with its corresponding caption and image while distinguishing it from its cultural counterpart.

To achieve this, we propose **CultureCLIP**, a contrastive learning framework that jointly embeds concepts, captions, and images into a shared semantic space. As illustrated in Figure 1, we encourage semantic **attraction** within each triplet and **repulsion** across its cultural counterpart. In addition, the overall architecture of our pipeline, including this learning scheme, is shown in Figure 2 (Right). We employ a shared text encoder for both concepts and captions to preserve the original alignment between images and captions while aligning concepts with images. This shared encoder design ensures that fine-tuning for cultural distinctions does not degrade the model's general cross-modal alignment ability. Our overall training objective is defined as:

$$\mathcal{L}_{\text{CultureCLIP}} = \lambda_c \cdot \mathcal{L}_{\text{concept}} + \lambda_t \cdot \mathcal{L}_{\text{caption}}, \tag{7}$$

where λ_c and λ_t balance the contributions of concept-level and caption-level objectives. Both objectives are symmetrically formulated to promote attraction within positive triplets and repulsion from their cultural counterparts:

$$\mathcal{L}_{\text{caption}} = \mathcal{L}_{\text{NegCLIP}}(I^+, T^+, T^-) + \mathcal{L}_{\text{NegCLIP}}(I^-, T^-, T^+), \tag{8}$$

$$\mathcal{L}_{\text{concept}} = \mathcal{L}_{\text{NegCLIP}}(I^+, C^+, C^-) + \mathcal{L}_{\text{NegCLIP}}(I^-, C^-, C^+), \tag{9}$$

This structure jointly anchors abstract cultural concepts to specific visual-textual cues, enhancing the model's ability to differentiate subtle cultural semantics. To further preserve the original model's generalization ability, we apply the parameter-efficient LoRA method for fine-tuning on this loss, rather than directly training both the visual and text encoders.

5 Experiment

5.1 Experimental Setup

Model Choices and Prompting Strategies In this work, we leverage Qwen2.5-VL (Bai et al., 2025) as our LLM/VLM (note that the specific choice of model is not the primary focus of this paper). Although more powerful LLMs could potentially further improve data quality, we consider Qwen2.5-VL to strike a good balance between performance and cost, making it a practical choice for our experiments. Given the significant impact that prompt design has on LLM performance (Brown et al., 2020), we meticulously craft prompt templates for each LLM in the pipeline, employing a few-shot learning approach. This includes presenting a set of example input-output pairs to guide the model in various tasks. All prompts used in this study are provided in Appendix F. For the text-to-image generation task, we use Stable Diffusion 3.5 (Rombach et al., 2022) as the model for image synthesis.

Benchmarks We evaluate the models on both culture-specific and culture-agnostic tasks. To assess cultural understanding, we adapt three benchmarks—GlobalRG-Grounding, GlobalRG-Retrieval (Bhatia et al., 2024), and CROPE (Nikandrou et al., 2024)—into statement-ranking tasks suitable for CLIP-based models. In each task, the model must select the most semantically accurate description for a given image from a set of culturally grounded statements, thereby testing its ability to capture fine-grained, culture-specific visual cues (reported as *Accuracy*). To assess general vision-language capabilities, we further evaluate the models on MS COCO (Lin et al., 2015) and Flickr30k (Plummer et al., 2016), reporting the average *Recall@5* for both image-to-text and text-to-image retrieval tasks. Details of the benchmarks and additional evaluation results on widely used image classification datasets are provided in Appendix D.

Baseline We first evaluate the performance of CLIP (Radford et al., 2021), NegCLIP (Yuksekgonul et al., 2022), and TripletCLIP (Patel et al., 2024) on our tasks to establish a baseline. Building on this, we introduce CLIP++, NegCLIP++, and TripletCLIP++ as enhanced baselines. In these ++ versions, we train the base CLIP model using our own dataset, aligning the synthetic images with contextualized caption-image pairs, without incorporating the concept.

Implementation Details All fine-tuned CLIP models use ViT-B/32 (Dosovitskiy et al., 2021) as the image encoder $\mathcal{F}$ and the default CLIP text encoder (Radford et al., 2021) as the text encoder $\mathcal{G}$. We fine-tune the model using a parameter-efficient method, LoRA (Hu et al., 2021), during which $\mathcal{F}$ and $\mathcal{G}$ are frozen, and additional LoRA parameters for these two encoders are applied and trained for 10 epochs with a global batch size of 2048, a learning rate of 3×10^{-6}, weight decay of 0.1, and a cosine learning rate schedule. We employ LoRA primarily to maintain the model's general capability rather than to reduce memory requirements. In our preliminary experiments with full parameter fine-tuning, we observed significant performance degradation on both culture-specific and culture-agnostic benchmarks, likely due to the distribution gap between the general data used in pretraining and the culture-specific data used during fine-tuning, which disrupted knowledge the model had already acquired. Before inference, LoRA parameters are merged with the backbone transformers, ensuring both efficiency and the preservation of zero-shot transfer capabilities. All models are fine-tuned on 4 Nvidia H20 GPUs using the official Hugging Face Transformers codebase (Wolf et al., 2020).

5.2 Main Results

Evaluation on Culture-Specific Tasks As summarized in Table 2, CultureCLIP significantly outperforms all baseline models on cultural benchmarks, achieving a 5.49% improvement over CLIP on GlobalRG-G, thus demonstrating strong fine-grained cultural understanding. Directly fine-tuning CLIP on our cultural dataset (CLIP++) leads to a substantial performance drop of 17.93%, highlighting the limitations of naive fine-tuning without explicit negative samples or concept-level alignment. In contrast, our enhanced variants NegCLIP++ and TripletCLIP++ (which incorporate hard negatives) achieve improvements of 2.34% and 0.80% over CLIP, respectively. When further combined with concept-level alignment in CultureCLIP, we observe a large net improvement, emphasizing the effectiveness of our design in leveraging hard negatives and concept supervision for cultural feature discrimination. Similar trends are observed for GlobalRG-R and CROPE. We note that NegCLIP and TripletCLIP are pre-trained based on much smaller datasets (e.g., CC3M, CC12M) and do not include culturally relevant data, resulting in substantially lower performance on general benchmarks. For fairness, we do not consider them direct baselines in the cultural evaluation but include them in the table for completeness. All models in Table 2 are trained on the same unfiltered 100k synthetic dataset using LoRA with rank 4 to ensure a fair comparison.

Evaluation on Culture-Agnostic Tasks On general vision-language tasks, CultureCLIP maintains strong performance and even slightly improves over the baseline, with gains of 0.90% on MS COCO and 0.30% on Flickr30k. This indicates that cultural fine-tuning does not compromise, and may even enhance, general retrieval capabilities.

5.3 Ablations

Which contributes more to the model and alignment, caption or concept? As shown in Table 3, concepts act as abstract semantic anchors, providing stronger cultural alignment capacity and enabling the model to distinguish subtle differences. Captions refine the model's understanding of specific details but are less critical for recognizing cultural nuances.

Can a higher-quality cultural dataset improve performance? As shown in Table 4, using quality-filtered data ("+QF"), which contains 73.8k high-quality samples after image filtering, consis-

Table 2: Experimental results on culture-specific and culture-agnostic tasks. All models are trained on the same unfiltered 100k dataset using LoRA with rank 4. Best scores are in **bold**. Second best scores are underlined.

Methods	Neg	Con	Culture-Specific Tasks			Culture-Agnostic Tasks	
			GlobalRG-G	GlobalRG-R	CROPE	MS COCO	Flickr30k
CLIP	×	×	63.98	78.22	74.69	65.40	89.0
NegCLIP	✓	×	-	-	-	6.50	2.70
TripletCLIP	✓	×	-	-	-	10.80	22.00
CLIP++ (ours)	×	×	46.05	49.98	73.62	28.80	50.80
NegCLIP++ (ours)	✓	×	66.32	78.41	79.25	65.50	89.20
TripletCLIP++ (ours)	✓	×	64.78	78.45	**79.25**	65.50	**89.30**
CultureCLIP (ours)	✓	✓	**69.47**	**78.60**	78.84	**66.30**	**89.30**

Table 3: Ablation study on loss configurations. All models are trained on the same unfiltered 100k dataset using LoRA (rank 4) to preserve general multimodal alignment capabilities.

Configuration	λ(cap/con)	Culture-Specific Tasks			Culture-Agnostic Tasks	
		GlobalRG-G	GlobalRG-R	CROPE	MS COCO	Flickr30k
Single Branch (No Negative)						
Caption-only w/o neg	1.0 / –	66.95	77.43	79.19	65.60	89.20
Concept-only w/o neg	– / 1.0	64.24	77.70	79.19	65.50	89.00
Single Branch (With Negative)						
Caption-only w/ neg	1.0 / –	66.27	77.53	79.25	65.50	89.20
Concept-only w/ neg	– / 1.0	65.83	77.29	79.19	65.60	88.90
Mixed Branches						
Cap (w/o neg) + Con (w/o neg)	0.5 / 0.5	67.29	77.27	79.25	65.50	89.20
Cap (w/ neg) + Con (w/o neg)	0.5 / 0.5	67.12	78.70	79.25	65.60	89.30
Cap (w/o neg) + Con (w/ neg)	0.5 / 0.5	68.81	76.87	79.19	65.50	89.20
Full (Both with Negative)						
Both w/ neg (Ours)	0.7 / 0.3	65.93	**78.80**	**79.37**	**66.80**	**89.50**
Both w/ neg (Ours)	0.5 / 0.5	67.12	78.25	78.60	66.30	89.20
Both w/ neg (Ours)	0.3 / 0.7	**69.47**	78.60	78.84	66.10	89.30

tently improves performance despite using fewer samples compared to the full 100k unfiltered set. For example, comparing Config 5 (+QF, LoRA r=4) to Config 3 (LoRA r=4), and Config 6 (+QF, LoRA r=8) to Config 4 (LoRA r=8), we observe clear performance gains across all cultural benchmarks. This underscores the effectiveness of our data curation pipeline in providing cleaner and more informative supervision for fine-grained cultural alignment.

What role does LoRA play in adapting a pretrained model to downstream tasks? Our ablation results in Table 4 show that without LoRA (e.g., Configs 1 and 2), directly fine-tuning a pretrained model on a domain-specific cultural dataset leads to a substantial performance drop—specifically, 22.52% and 26.03% lower compared to Configs 3 and 5, respectively. This suggests that, without LoRA, the model struggles to effectively absorb specialized supervision, resulting in a significant loss of generalization ability. In contrast, when LoRA is applied (Configs 3–6), the model not only achieves improved performance on cultural benchmarks but also maintains its capabilities on general tasks. These findings highlight LoRA's essential role in mitigating catastrophic forgetting: it enables the model to flexibly adapt to new cultural signals while preserving the broad vision-language alignment learned during pretraining, ensuring robustness across both domain-specific and general scenarios.

Table 4: Ablation study on quality filtering (QF) and LoRA. "+QF" uses the 73.8k filtered samples; otherwise, the full 100k dataset is used.

Configuration	QF	LoRA Rank	GlobalRG-G	GlobalRG-R	CROPE
Baseline	×	-	46.95	49.98	73.62
+ QF	✓	-	43.64	50.68	74.33
+ LoRA (r=4)	×	4	69.47	**78.60**	78.84
+ LoRA (r=8)	×	8	65.29	78.03	**79.37**
+ QF + LoRA (r=4)	✓	4	**69.67**	**78.60**	79.21
+ QF + LoRA (r=8)	✓	8	66.48	78.21	**79.37**

6 Conclusion

In this paper, we introduce CulTwin, a high-quality synthetic dataset of paired concept-caption-image triplets verified by humans, where captions are enriched with cultural background knowledge using a vision-language model, and images are generated by a text-to-image model to reflect fine-grained visual features. Building on CulTwin, we propose CultureCLIP, a novel contrastive learning framework that jointly aligns cultural concepts, captions, and images in a shared embedding space. Our experiments demonstrate that CultureCLIP surpasses baseline models on culture-specific benchmarks, achieving a 5.49% improvement while simultaneously showing performance gains rather than degradation on culture-agnostic benchmarks. These results underscore the effectiveness of our synthetic dataset and training methodology in capturing nuanced cultural distinctions while preserving and even enhancing the model's generalization capabilities across broader contexts.

7 Limitations and Future Work

While CultureCLIP significantly improves fine-grained cultural understanding, several limitations remain. Both CLIP and CultureCLIP still struggle with cases where the visual distinction is highly abstract or stylistic (see error case analysis in Appendix E). In addition, our current pipeline, for practical considerations, adopts Qwen2.5-VL as an MLLM-as-a-Judge to assess cultural relevance through multidimensional scoring and to guide filtering. However, compared to more advanced models such as GPT-4o, this choice may introduce biases, lead to misjudgments, or lack interpretability. Furthermore, despite the diversity of CulTwin, it is fundamentally a fully synthetic dataset, and there may still exist a distributional gap between synthetic and real-world images. Looking ahead, future work may explore several directions: **(1) Improving abstract visual reasoning**, by enhancing the model's capacity to recognize subtle visual cues, such as artistic styles or symbolic meanings; **(2) Enhancing the robustness of cultural understanding**, by mitigating vulnerability to visual variations and ensuring stable performance across diverse conditions (Fan et al., 2025); **(3) Developing interpretable assessment modules**, to enable more robust and transparent cultural data evaluation beyond the current Qwen2.5-VL-based judge; and **(4) Bridging the synthetic-real domain gap**, by integrating real and synthetic data or applying domain adaptation strategies to improve generalization and visual grounding.

References

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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features, 2023. URL https://arxiv.org/abs/2308.11485.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark, 2024. URL https://arxiv.org/abs/2402.04788.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. Unveiling the lack of lvlm robustness to fundamental visual variations: Why and path forward, 2025. URL https://arxiv.org/abs/2504.16727.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip), 2022. URL https://arxiv.org/abs/2205.01397.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024.

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R. Fung. Mmboundary: Advancing mllm knowledge boundary awareness through reasoning step confidence calibration, 2025. URL https://arxiv.org/abs/2505.23224.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL https://arxiv.org/abs/2102.05918.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. Jina clip: Your clip model is also your text retriever, 2024. URL https://arxiv.org/abs/2405.20204.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *arXiv preprint arXiv:2408.07246*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*, 2025.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, et al. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip, 2023. URL https://arxiv.org/abs/2208.05516.

Malvina Nikandrou, Georgios Pantazopoulos, Nikolas Vitsakis, Ioannis Konstas, and Alessandro Suglia. Crope: Evaluating in-context adaptation of vision and language models to culture-specific concepts. *arXiv preprint arXiv:2410.15453*, 2024.

Maitreya Patel, Naga Sai Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, et al. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in Neural Information Processing Systems*, 37:32731–32760, 2024.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl, 2025. URL https://arxiv.org/abs/2503.07536.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. URL https://arxiv.org/abs/1505.04870.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024. URL https://arxiv.org/abs/2406.05967.

Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions, 2024. URL https://arxiv.org/abs/2401.02460.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks?, 2021. URL https://arxiv.org/abs/2107.06383.

Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6088–6100, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.421. URL https://aclanthology.org/2022.acl-long.421/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pp. 310–325. Springer, 2024a.

Gengyuan Zhang, Yurui Zhang, Kerui Zhang, and Volker Tresp. Can vision-language models be a good guesser? exploring vlms for times and location reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 636–645, 2024b.

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R. Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues, 2025. URL https://arxiv.org/abs/2502.12084.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13774–13784, 2024c.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.05379.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, July 2022. ISSN 1573-1405. doi: 10.1007/s11263-022-01653-1. URL http://dx.doi.org/10.1007/s11263-022-01653-1.

A Country List and Cultural Taxonomy

In this study, the list of country names is based on data from the GeoNames database: GeoNames.org.

The cultural taxonomy includes the following categories, each representing a significant aspect of cultural identity:

- **Cuisine**: Refers to the foods, culinary practices, and cooking methods that are unique to specific regions or cultures. This includes iconic dishes, preparation techniques, and the cultural background behind eating habits, as well as the importance of food in social and religious practices.
- **Clothing**: Encompasses traditional garments, accessories, and adornments from various cultures. It includes not only clothing but also items like jewelry, headwear, and footwear that hold cultural significance, reflecting identity, status, and traditions.
- **Animal & Plants**: Describes the native species, both fauna and flora, that hold cultural importance. This category includes the use of animals and plants in mythology, cuisine, traditional medicine, and environmental practices, as well as their roles in folklore and symbolism.
- **Art**: Includes visual arts, sculptures, and other forms of artistic expression that represent a culture's aesthetic and artistic heritage. This encompasses paintings, sculptures, performance arts, and crafts that reflect the identity, beliefs, and historical evolution of a community.
- **Architecture**: Refers to the design, style, and structures built by a particular culture. This includes traditional houses, temples, monuments, and public buildings that showcase the engineering, material use, and aesthetic values of the culture.
- **Daily Life**: Covers the everyday activities, routines, and practices that define how people in a particular culture live. This includes family roles, work habits, and leisure activities, as well as practices around health, education, and community.
- **Symbol**: Involves the symbols, logos, and imagery that carry cultural meaning. This category includes national flags, religious icons, mythological figures, and colors that convey beliefs, values, and identity in various contexts.
- **Festival**: Encompasses cultural festivals, holidays, and ceremonies, along with the associated customs, rituals, and practices. Examples include events like Chinese New Year, Diwali, and Christmas, each rich in traditions, foods, and rituals that symbolize community and heritage.

B CulTwin Details

We collected a total of 99,996 *Twin Cards*, each consisting of two concept–caption–image triplets with culturally distinct negatives. After applying the image quality filtering described in Section 3.3, 73,823 high-quality samples were retained—a scale significantly larger than existing cultural benchmarks such as CROPE (approximately 1k samples). Each concept is represented as a single word, and the corresponding captions have an average length of 14.55 words. All images are synthetically generated from these captions using Stable Diffusion 3.5 Large Turbo, with an efficient generation throughput of approximately 3,000 images per hour on a single H20 GPU. Additional examples of *Twin Cards* are illustrated in Figure 4.

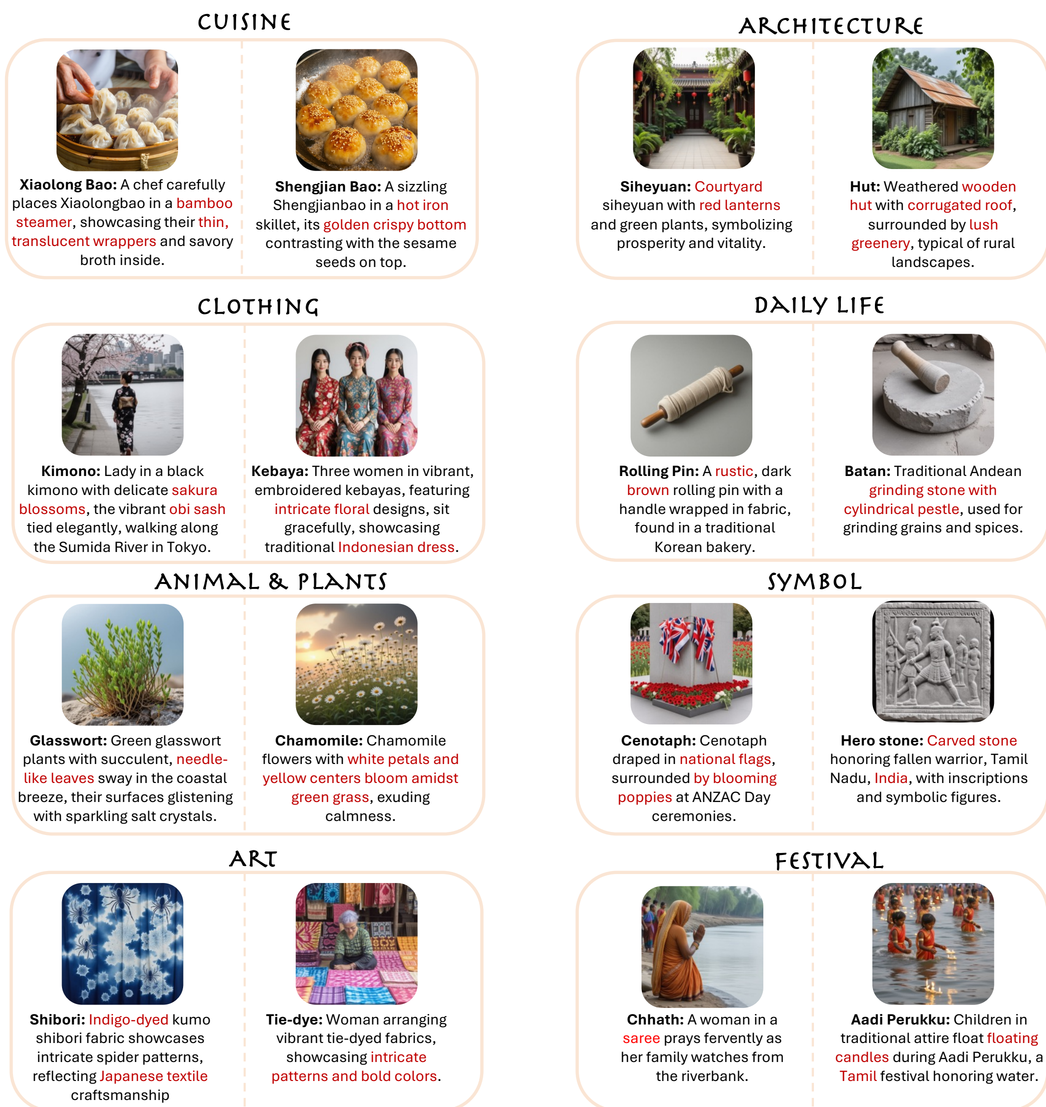

Figure 4: Additional *Twin Card* examples showcasing diverse cultural concepts beyond the main text examples, further illustrating fine-grained cultural distinctions and visual similarity.

C CultureCLIP Pseudocode

```
def cultureclip_loss(pos_concept_embs, pos_caption_embs,
    pos_img_embs,
                      neg_concept_embs, neg_caption_embs,
                          neg_img_embs,
                      logit_scale, lambda_caption=0.5,
                          lambda_concept=0.5):
    """
    Full CultureCLIP Loss: Caption's negclip + Concept's negclip
    Aligns concepts with their corresponding captions and images,
    while distinguishing them from their culturally opposite
        counterparts.
    """

    # Calculate Caption's NegCLIP loss
    caption_loss_val = (
        negclip_loss(pos_img_embs, pos_caption_embs,
            neg_caption_embs, logit_scale) +
        negclip_loss(neg_img_embs, neg_caption_embs,
            pos_caption_embs, logit_scale)
    )

    # Calculate Concept's NegCLIP loss
    concept_loss_val = (
        negclip_loss(pos_img_embs, pos_concept_embs,
            neg_concept_embs, logit_scale) +
        negclip_loss(neg_img_embs, neg_concept_embs,
            pos_concept_embs, logit_scale)
    )

    # Total Loss
    total_loss = lambda_caption * caption_loss_val + lambda_concept
         * concept_loss_val

    return total_loss
```

D Benchmark Details and Additional Results

In this section, we provide detailed descriptions of the benchmarks used to evaluate our models, covering both culture-specific and culture-agnostic tasks.

Culture-Specific Tasks

- **GlobalRG-Grounding** (Bhatia et al., 2024): Each data point consists of an image, a concept, and a country. We generate four statement-based options for the model to choose from, such as ”The item in the picture is {concept} in {country}.” The correct option corresponds to the appropriate concept for that country, while incorrect options are created by selecting a concept from the same country that does not match the image.
- **GlobalRG-Retrieval** (Bhatia et al., 2024): Each data point consists of an image, a category, and a country, without specifying a particular concept. The task focuses on identifying the correct country for the depicted category. Options are phrased as ”The picture depicts a kind of {category} in {country},” with incorrect options generated by randomly substituting the country.
- **CROPE** (Nikandrou et al., 2024): The original dataset asks whether the image displays the defined concept (”yes” or ”no”). We filter out ”no” cases where the question concept and definition concept differ, and reformulate the task as a two-choice classification: ”There is {question concept} in the image” or ”There is {definition concept} in the image,” requiring the model to correctly identify the depicted concept.

Culture-Agnostic Tasks

- **MS COCO** (Lin et al., 2015): This large-scale dataset comprises over 330,000 images annotated with approximately five captions each. We evaluate using bidirectional retrieval metrics, specifically *Text2Image* and *Image2Text* Recall@5, and report their arithmetic mean as the final score.
- **Flickr30k** (Plummer et al., 2016): This dataset contains 31,000 images primarily focused on human activities, each paired with five descriptive captions. Similar to MS COCO, we adopt bidirectional retrieval metrics to assess cross-modal alignment.
- **More General Image Classification Benchmarks**: To further assess generalization capabilities, we evaluate our models on several widely used image classification benchmarks, including **FER2013**, **ImageNet-1k**, **ImageNet-A**, **ImageNet-O**, **ImageNet-R**, **VOC2007**, **CIFAR-10**, and **CIFAR-100**. These datasets collectively cover a broad spectrum of visual tasks, such as facial expression recognition, large-scale object classification, out-of-distribution robustness, multi-label classification, and both coarse- and fine-grained category recognition. Empirical results demonstrate that CultureCLIP consistently maintains, and in some cases slightly improves, performance on these benchmarks. This suggests that the incorporation of cultural fine-tuning does not compromise general vision-language alignment or classification capabilities but rather enhances overall robustness and versatility. A comprehensive summary of these results, including Top-1 Accuracy (*Acc1*), Top-5 Accuracy (*Acc5*), and Mean Per-Class Recall (*MPCR*), is provided in Table 5.

Table 5: Performance on general image classification benchmarks (%). We report Top-1 Accuracy (Acc1), Top-5 Accuracy (Acc5), and Mean Per-Class Recall (MPCR) across various datasets. CultureCLIP maintains or slightly improves general performance despite additional cultural training.

Model	FER2013	ImageNet-1k	ImageNet-A	ImageNet-O	ImageNet-R	VOC2007	CIFAR-10	CIFAR-100
Top-1 Accuracy (Acc1)								
OpenAI CLIP	41.22	63.37	31.51	47.55	69.31	76.45	89.77	64.24
Caption (w/o neg)	41.57	63.37	**31.61**	**47.80**	69.28	**76.52**	89.84	**64.48**
Caption (w/o neg) + Concept (w/o neg)	41.20	63.35	31.59	47.65	**69.33**	76.42	**89.86**	64.40
Caption (w/ neg)	**41.59**	**63.39**	**31.61**	47.65	69.26	**76.52**	89.80	64.45
CultureCLIP	41.26	63.37	31.60	47.65	69.31	76.48	89.83	64.47
Top-5 Accuracy (Acc5)								
OpenAI CLIP	94.78	88.82	64.23	**78.30**	88.81	95.93	99.61	88.78
Caption (w/o neg)	94.68	88.82	64.31	78.10	88.80	**95.98**	99.62	**88.91**
Caption (w/o neg) + Concept (w/o neg)	**94.82**	**88.83**	64.20	78.25	88.00	95.94	**99.63**	88.84
Caption (w/ neg)	94.69	88.82	**64.33**	78.10	88.79	95.96	**99.63**	88.87
CultureCLIP	94.80	88.80	64.29	78.10	**88.85**	95.95	**99.63**	88.85
Mean Per-Class Recall (MPCR)								
OpenAI CLIP	36.10	63.36	32.63	48.85	67.92	80.59	89.83	64.21
Caption (w/o neg)	36.40	63.36	32.75	**49.09**	67.92	80.60	89.86	**64.44**
Caption (w/o neg) + Concept (w/o neg)	36.00	63.37	32.72	49.01	67.94	**80.64**	**89.87**	64.40
Caption (w/ neg)	**36.42**	**63.39**	32.72	49.04	67.87	80.58	89.83	64.41
CultureCLIP	36.21	63.37	**32.77**	49.00	**67.96**	80.60	89.85	64.41

E Error Case Analysis

Distinguishing between Gongbi and Xieyi styles One illustrative failure case involves differentiating between two classic Chinese painting styles: *gongbi* and *xieyi*. The *gongbi* style is known for its meticulous brushwork, fine lines, and realistic details, often used to depict flowers, birds, and other subjects in a highly controlled and precise manner. In contrast, the *xieyi* style (literally "writing ideas") emphasizes freehand expression, bold strokes, and abstract or suggestive forms rather than realistic details. In this example, both CLIP and CultureCLIP misclassified an input *gongbi* painting as *xieyi*. However, CultureCLIP assigned a lower confidence to the incorrect label (72% vs. CLIP's 78%), indicating a modest calibration improvement. This suggests that while our model still struggles with highly abstract stylistic distinctions, it demonstrates better uncertainty awareness compared to the original CLIP, which is a step toward more nuanced cultural reasoning.

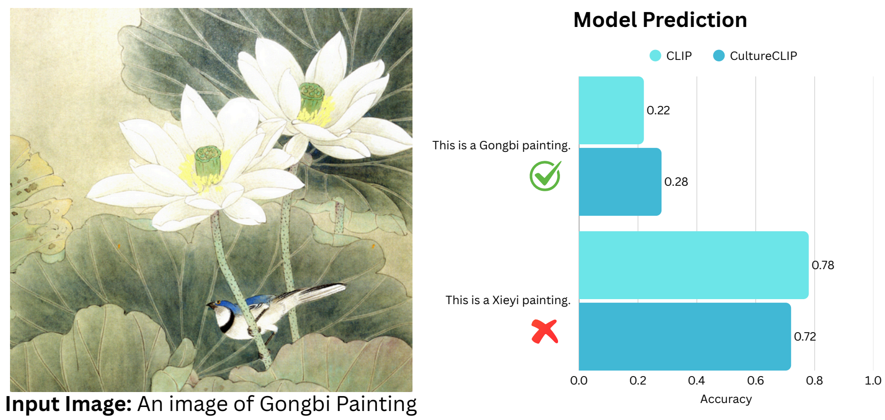

Figure 5: Example failure case comparing CLIP and CultureCLIP on a *gongbi* painting. While both models misclassified it as *xieyi*, CultureCLIP exhibited lower confidence in its wrong prediction, suggesting better calibration.

F Prompt Engineering

Bottom-up Filtering Prompt for Qwen2.5-VL

```
Please determine whether the concept '{title}' clearly and
unambiguously belongs to one of the eight cultural categories (
Cuisine, Clothing, Animals & Plants, Art, Architecture, Daily Life,
 Symbol, or Festival). If the concept is only loosely related,
culturally ambiguous, or does not strongly align with any of the
categories, please select 'A' to ensure strict filtering. The
concept does not clearly belong to any of the eight categories. The
 concept is clearly and directly related to one of the eight
categories. Answer only with 'A' or 'B'.

Output the result in JSON format as follows:
{
    "concept_type": "A" or "B",
}
```

Figure 6: Bottom-up filtering prompt for Qwen2.5-VL.

Bottom-up Classification Prompt for Qwen2.5-VL

```
Given a cultural concept, definition, definition caption, and
definition image, your task is to classify and extract the
following information in English:
    - Country: The country or region associated with the concept.
    - Cultural Category: IMPORTANT - You MUST choose exactly ONE
    category from these eight predefined categories:
        * Cuisine
        * Clothing
        * Animal & Plants
        * Art
        * Architecture
        * Daily Life
        * Symbol
        * Festival
        Do not use any other categories. If the concept doesn't
        clearly fit into one of these categories, choose the
        closest match.
    - Context: A short 20-word description in English that provides
     insight into the cultural and functional use of the concept.
    - Key Visual Features: The visual features that distinguish the
     concept (e.g., shape, material, color, size) described in
    English.

Examples:
    Input:
        Concept: Kimono
        Definition: A traditional Japanese garment with long, wide
        sleeves and an obi sash, worn for formal occasions and
        festivals.
        Definition Caption: A colorful kimono with crane patterns,
        displayed at Kyoto's textile museum.
        Definition Image: example_image.jpg
    Output:
        Country: Japan
        Category: Clothing
        Context: A traditional Japanese garment with long, wide
        sleeves and an obi sash, worn for formal occasions and
        festivals.
        Key Visual Features: Long, wide sleeves, obi sash, colorful
         patterns, traditional Japanese style.

Current Task:
    Concept: {concept}
    Definition: {definition}
    Definition Caption: {caption}
    Definition Image: {image_url}
    Generate the Country, Cultural Category (MUST be one of the
    eight predefined categories), Context, and Key Visual Features
    in English:
```

Figure 7: Bottom-up classification prompt for Qwen2.5-VL.

Top-down Generation Prompt for Qwen2.5-VL

```
Given a country and cultural category, your task is to generate the
 following:
    - Concept: A cultural concept that fits the provided country
    and category.
    - Context: A short 20-word description that provides insight
    into the cultural and functional use of the concept.
    - Key Visual Features: The visual features that distinguish the
     concept (e.g., shape, material, color, size).

Examples:
    Input:
        Country: China
        Cultural Category: Food
    Output:
        Concept: Mantou
        Context: Steamed wheat bun symbolizing prosperity and
        wisdom, commonly served during festivals and family
        gatherings.
        Key Visual Features: Pillowy white appearance, round or
        rectangular shape, smooth surface, typically palm-sized.
    Input:
        Country: Japan
        Cultural Category: Art
    Output:
        Concept: Ukiyo-e
        Context: Traditional woodblock prints depicting scenes from
         everyday life, nature, and historical events.
        Key Visual Features: Flat color blocks, bold outlines,
        vibrant pigments, rectangular format, detailed patterns.

Current Task:
    Country: {country}
    Cultural Category: {category}
    Generate the concept, context, and key visual features:
```

Figure 8: Top-down generation prompt for Qwen2.5-VL.

Twin Matching Prompt for Qwen2.5-VL

```
Given a cultural concept from a certain category, its context, and
its visual features, your task is to generate the following:
    - New Concept: A visually similar but culturally different
    concept from the same category.
    - New Context: A short 20-word description in English that
    provides insight into the cultural and functional use of the
    generated concept.
    - New Key Visual Features: The visual features that distinguish
     the concept (e.g., shape, material, color, size) from the
    original concept.

Examples:
    Input:
        Category: Art
        Concept: Erhu
        Context: A two-stringed Chinese musical instrument played
        with a bow, often used in traditional Chinese music.
        Key Visual Features: Two strings, a bow, a wooden body, and
         a horsehair bow.
    Output:
        New Concept: Guzheng
        New Context: A Chinese zither-like instrument with a large,
         rectangular wooden body and multiple strings, played with
        plucking.
        New Key Visual Features: A large, rectangular wooden body,
        multiple strings, and a plucking mechanism.

Current Task:
    Category: {category}
    Concept: {concept}
    Context: {context}
    Visual Features: {visual_features}
    Generate a new concept, context, and visual features:
```

Figure 9: Twin matching prompt for Qwen2.5-VL.

Diverse Caption Generation Prompt for Qwen2.5-VL

```
Given a cultural concept, context, and key visual features, your
task is to generate 10 different captions that describe the concept
 in various scenarios while preserving its cultural significance.
The captions must reflect different styles or settings, but they
should all clearly include the key visual features. Follow these
guidelines:
    - Emphasize the key visual differences in the scenes (e.g.,
    shape, size, cooking method, setting).
    - Retain the cultural or functional context (e.g., everyday use
    , ceremonial purpose, tradition).
    - Ensure the differentiating features (e.g., shape, texture,
    size, material, use) are clearly reflected in each caption.
    - Each caption should be under 15 words.
    - Each caption should be unique, showing different perspectives
     or settings, but should always include the key visual features
    .

Examples:
    Input:
        Concept: Xiaolongbao (Soup Dumplings)
        Context: Steamed soup dumplings with delicate, thin
        wrappers, filled with savory broth and pork, typically
        steamed in bamboo baskets.
        Key Visual Features: Delicate, thin wrappers and steamed in
         bamboo baskets.
    Output:
        1. A chef carefully places Xiaolongbao in a bamboo steamer,
         showcasing their thin, translucent wrappers and savory
        broth inside.
        2. A steaming bamboo basket of Xiaolongbao, delicate
        wrappers holding savory broth, served in an elegant
        Shanghai restaurant.
        3. Steamed xiaolongbao resting in bamboo baskets, ready to
        be served during a family meal.
        4. Crispy fried xiaolongbao, golden-brown and served with
        dipping sauce, sitting in bamboo baskets.
        5. Miniature xiaolongbao filled with crab roe, elegantly
        presented in bamboo baskets at a Cantonese restaurant.
        6. Steaming xiaolongbao with delicate skin, served in
        bamboo baskets during a traditional Chinese New Year meal.
        7. Bamboo-steamed xiaolongbao, filled with savory broth,
        served alongside hot tea in a Beijing teahouse.
        8. Translucent, plump xiaolongbao, freshly steamed in
        bamboo baskets for a cozy brunch setting.
        9. Steamed xiaolongbao with pork filling, served in bamboo
        baskets with chili oil at a street food stall.
        10. Elegant xiaolongbao, arranged in bamboo baskets,
        presented at a lavish festive feast.

Current Task:
    Concept: {concept}
    Context: {context}
    Key Visual Features: {visual_features}
    Generate 10 different captions, each reflecting a different
    style or scene, but all incorporating the key visual features:
```

Figure 10: Diverse caption generation prompt for Qwen2.5-VL.

Prompt for Quality Evaluation on Authenticity

```
Please analyze this image and rate its authenticity on a scale of 1
 to 5. You can refer to the context to help you make the decision.
Focus on whether the concept shown is realistic and follows common
sense. Consider:
    - Are all elements anatomically and physically correct?
    - Does everything look natural and possible in the real world?
    - Are there any unrealistic or deformed features?

Examples:
    Input:
        Concept: Erhu
        Context: The Erhu is a traditional Chinese musical
        instrument played with a bow. It consists of a wooden body
        and two strings, and is known for its expressive and
        resonant sound. Often used in Chinese classical and folk
        music, it is typically performed in both solo and ensemble
        settings.
        Image: An image showing a person playing Erhu with three
        hands in mid air
    Output: 1
    Input:
        Concept: Erhu
        Context: The Erhu is a traditional Chinese musical
        instrument played with a bow. It consists of a wooden body
        and two strings, and is known for its expressive and
        resonant sound. Often used in Chinese classical and folk
        music, it is typically performed in both solo and ensemble
        settings.
        Image: An image showing a person playing Erhu with two
        hands sitting on the chair;
    Output: 5

Now, give you the image, concept and its corresponding context:
    Concept: {concept}
    Context: {context}
    Output only the score:
```

Figure 11: Prompt for quality evaluation on authenticity.

Prompt for Quality Evaluation on Consistency

```
Please analyze this image and rate its consistency with the concept
 on a scale of 1 to 5. You can refer to the context to help you
make the decision. Focus on whether the image accurately depicts
the specified concept without showing wrong concepts. Consider:
    - Does the image show exactly the concept mentioned?
    - Are there any mismatched or wrong elements?
    - Is the concept clearly and accurately represented?

Examples:
    Input:
        Concept: Tang Sancai
        Context: A type of Chinese glazed pottery from the Tang
        Dynasty, known for its colored glazes (green, yellow, and
        brown) often used for tomb figurines and decorative pieces.
        Image: An image showing a blue-and-white porcelain bowl
    Output: 1
    Input:
        Concept: Tang Sancai
        Context: A type of Chinese glazed pottery from the Tang
        Dynasty, known for its colored glazes (green, yellow, and
        brown) often used for tomb figurines and decorative pieces.
        Image: An image showing exactly a Tang Sancai horse
    Output: 5

Now, give you the image, concept and its corresponding context:
    Concept: {concept}
    Context: {context}
    Output only the score:
```

Figure 12: Prompt for quality evaluation on consistency.

Prompt for Quality Evaluation on Cultural Fidelity

```
Please analyze this image and rate its cultural fidelity on a scale
 of 1 to 5. Focus on whether the cultural elements are accurate and
 appropriate for the specific context. Consider:
    - Are all cultural elements accurate for this context?
    - Are there any mixed or incorrect cultural elements?
    - Does everything align with the cultural background specified?

Examples:
    Input:
        Concept: Mexican Day of the Dead
        Context: A Mexican holiday celebrating deceased loved ones,
         featuring marigold flowers, sugar skulls, and candle
        altars.
        Image: An image showing marigold flowers, skull paintings,
        and candle altars.
    Output: 5
    Input:
        Concept: Mexican Day of the Dead
        Context: A Mexican holiday celebrating deceased loved ones,
         featuring marigold flowers, sugar skulls, and candle
        altars.
        Image: An image showing showing marigold flowers, skull
        paintings, and Chinese joss paper money.
    Output: 1

Now, give you the image, concept and its corresponding context:
    Concept: {concept}
    Context: {context}
    Image:{image}
    Output only the score:
```

Figure 13: Prompt for quality evaluation on cultural fidelity.