Beyond Static Symbols: Evidence-Grounded Visual Checklists for Cultural Activity Generation

Hongcan Deng[1,2,3], Bo Ma[1,2,3,*], Yating Yang[1,2,3,*], Rong Ma[1,2,3], Zhen Wang[1,2,3], Xi Zhou[1,2,3], Rui Dong[1,2,3], Turghun Osman[1,2,3]

[1]Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China

denghongcan24@mails.ucas.ac.cn, {mabo,yangyt,marong,wang_zhen,zhouxi,dongrui,turghun}@ms.xjb.ac.cn

Abstract

Text-to-image models can render recognizable cultural symbols, but cultural activity generation requires coordinated control over participants, actions, interactions, scenes, and culture-specific constraints. We propose CARE, an evidence-grounded visual checklist framework for controllable cultural activity image generation. Given an activity prompt, CARE builds source-traceable cultural evidence, converts it into an optimized prompt and fine-grained checklist, verifies generated images with concrete VLM-based visual checks, and uses failed checks for targeted prompt repair. We introduce CARE-1360, a dataset of 1,360 cultural activity prompts across eight countries, and evaluate generated images with a unified human protocol. Experiments show that CARE consistently improves over direct generation and retrieval-based prompt refinement baselines across cultural and visual quality dimensions. Results with both ZImage and Stable Diffusion XL suggest that evidence-grounded visual checklists provide an interpretable control mechanism that transfers across the tested generation backends.

Code and data: https://github.com/HeXCatalyst/CARE.

1 Introduction

Text-to-image models are increasingly capable of generating semantically relevant and visually appealing images from natural-language prompts (Ramesh et al., 2021; Rombach et al., 2022; Saharia et al., 2022). However, in culturally grounded scenarios, generating images that contain salient cultural symbols is not equivalent to generating culturally correct images. This distinction is particularly important for cultural activities, where an image must not only depict appropriate clothing, objects, and environments, but also correctly organize participant roles, core actions, human interactions, and spatiotemporal context, while avoiding cross-cultural mixing, stereotypical symbols, and overly staged expressions. Therefore, cultural activity generation is not merely a problem of aligning cultural nouns or visual attributes; rather, it is a structured generation-control problem involving objects, actions, scenes, and cultural constraints.

Existing work on culture-aware text-to-image generation has primarily focused on the visual grounding of cultural concepts, cultural objects, or culture-specific nouns (Cho et al., 2023; Liu et al., 2023; Basu et al., 2023; Kannen et al., 2024; Liu et al., 2024; Rege et al., 2025; Nayak et al., 2025). For example, Culture-TRIP (Jeong et al., 2025) retrieves cultural background knowledge and visual details, and iteratively optimizes prompts to improve image alignment for underrepresented culture nouns. Such methods mitigate representational gaps at the level of cultural entities. However, when the target shifts from "generating a cultural object" to "generating a cultural activity," models must also jointly organize roles, actions, interaction relations, and scene contexts, which cannot be fully captured by static noun grounding alone. For instance, consider an image of "a Chinese family sweeping an ancestral grave during the Qingming Festival." If the image merely contains objects such as incense, flowers, or a tombstone, but fails to clearly depict family members cleaning the grave, arranging offerings, and jointly performing the memorial activity in a quiet early-spring setting, it may still fail to

*Corresponding authors

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 15267–15294
October 24-29, 2026 ©2026 Association for Computational Linguistics

faithfully represent the cultural activity.

In this paper, we propose CARE, an evidence-driven closed-loop generation framework for cultural activity image generation. Its core idea is to transform an abstract activity description, before image generation, into visualizable, verifiable, and source-traceable cultural evidence, and to use this evidence as a shared basis for generation and evaluation. The system no longer relies on a model's implicit associations about cultural activities. Instead, it explicitly identifies the key objects, human actions, interaction relations, scene contexts, and cultural errors that should be avoided. After image generation, we reformulate cultural correctness assessment from a holistic judgment into fine-grained visual verification: rather than directly asking a VLM whether an image is culturally correct, the system checks concrete visual conditions such as core actions, participant relations, scene context, and cross-cultural mixing or stereotypical elements. When an image fails to satisfy these conditions, the system repairs the prompt only with respect to the failed evidence, and prioritizes subsequent candidates that resolve the original issues while preserving already satisfied conditions. In this way, cultural activity image generation is modeled as a diagnosable, traceable, and iteratively improvable process, rather than a one-shot prompt expansion procedure or a simple ranking of candidate images.

Our design is inspired by prior work on interpretable text-to-image evaluation and cultural activity assessment. TIFA (Hu et al., 2023) shows that fine-grained visual question answering provides more specific image-text faithfulness error localization than holistic scoring. CULTIVate (Malakouti et al., 2026) further argues that evaluating culturally grounded generation should not only focus on object-centric symbols such as food, clothing, or landmarks, but also examine observable dimensions in social activities, including backgrounds, interactions, hallucinations, and exaggerated expressions. Inspired by these insights, we decompose cultural activities into observable and verifiable visual elements for image evaluation, failure analysis, and subsequent repair. Accordingly, this paper does not aim to simply increase the number of cultural symbols in generated images. Instead, it seeks to better represent the key objects, human actions, interaction relations, and scene contexts involved in cultural activities, while reducing cross-cultural mixing and stereotypical expressions. By feeding evaluation results back into generation, the system can identify missing or incorrect visual elements and make targeted adjustments in subsequent candidates, turning cultural activity image generation into a more interpretable and traceable process of generation and repair.

Our contributions are threefold:

- We formalize cultural activity image generation as structured control over objects, actions, interactions, scenes, and cultural constraints, going beyond cultural nouns or static objects.
- We propose evidence-grounded visual checklists that turn activity descriptions into observable, verifiable, and source-traceable generation/evaluation criteria.
- We introduce checklist-guided closed-loop repair, using failed checks and linked evidence for targeted prompt revision, and validate it on ZImage and SDXL.

2 Related Work

2.1 Culture-Aware Text-to-Image Generation

Culture-aware T2I work has studied representation biases and culturally specific generation across cultural objects, regional attributes, cultural competence, and cultural expectations (Cho et al., 2023; Luccioni et al., 2023; Liu et al., 2023; Basu et al., 2023; Kannen et al., 2024; Liu et al., 2024; Rege et al., 2025; Nayak et al., 2025). Image transcreation further studies culturally adapting images while preserving meaning, showing that culturally relevant visual adaptation remains difficult for current generative pipelines (Khanuja et al., 2024). Recent VLM benchmarks also examine cultural visual reasoning and grounding, highlighting that recognizing culturally specific artifacts and grounding them spatially remain challenging (Satar et al., 2025). A related line of work shows that cultural failures in T2I are not limited to missing representation: models may also produce visual stereotypes, over-symbolized depictions, or narrow visual tropes for complex social concepts (Basu et al., 2023; Nayak et al., 2025). However, most prior work remains object- or attribute-centric, focusing on cues such as clothing, food, architecture, and landmarks; even Culture-TRIP (Jeong et al., 2025) uses retrieved knowledge mainly to refine prompts for underrepresented cultural nouns. CULTIVate (Malakouti et al., 2026) and CulturalFrames (Nayak et al., 2025) are closer to our setting

China / Spring Festival — "A Chinese family gathers around a table making jiaozi dumplings together on Spring Festival Eve, laughing as they prepare for the new year."



Base C-TRIP(3) **Ours**

Figure 1: Main qualitative comparison for a representative cultural activity prompt. Additional qualitative examples from Brazil, Iran, Germany, and South Africa are provided in Appendix G.

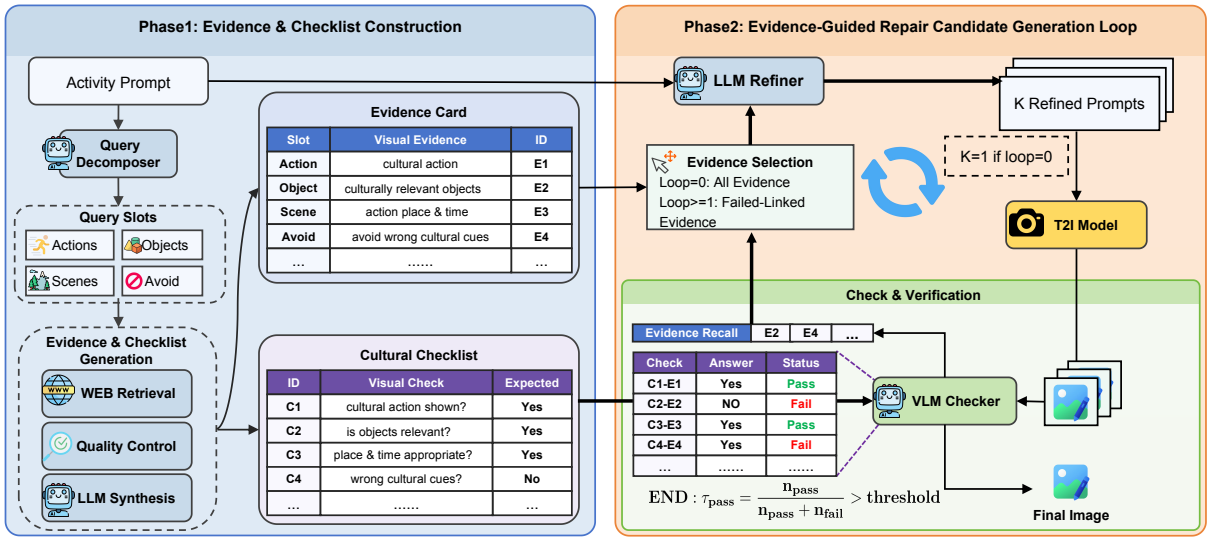


Figure 2: Overview of CARE. The system constructs an Evidence Card and checklist, verifies generated images with checklist-based VLM judgments, and uses failed checks to guide evidence-grounded prompt repair.

because they emphasize social activities and human cultural expectations, but CARE further uses cultural evidence as a shared control space for activity generation, checklist-based verification, and targeted repair.

2.2 Fine-Grained Evaluation for Text-to-Image Generation

T2I evaluation has shifted from holistic scores to fine-grained visual checks, including VQA-based error localization, low-redundancy atomic checks, and VLM-based evaluation metrics (Hu et al., 2023; Huang et al., 2023; Cho et al., 2024; Ku et al., 2024; Li et al., 2024). In culturally aware multimodal understanding, Seeing Culture provides culturally oriented visual reasoning questions and artifact grounding annotations for evaluating whether VLMs recognize and localize culturally specific visual elements (Satar et al., 2025). These methods improve interpretability for missing objects, incorrect relations, and semantic mismatches, but they mainly target general prompt faithfulness or image understanding rather than culturally grounded activity generation. CARE instead converts cultural evidence into observable checks over objects, actions, scenes, and avoid constraints, and uses failed checks for both error localization and prompt repair.

2.3 Prompt Refinement and Evidence-Grounded Generation

Prompt optimization and retrieval-augmented methods improve T2I generation by rewriting inputs, iteratively revising prompts with multimodal feedback, or supplementing models with external evidence (Cao et al., 2023; Yeh et al., 2026; Khan

et al., 2025; Lewis et al., 2020; Asai et al., 2024). In cultural T2I, Culture-TRIP retrieves cultural background and visual details to optimize prompts for cultural nouns (Jeong et al., 2025). Related work on cultural heritage retrieval also shows that LLM-augmented descriptions require explicit completeness and consistency validation against visual cues (Zhang et al., 2025a). CARE differs by organizing retrieved evidence into an Evidence Card and recalling it according to failed checklist items, so prompt repair targets concrete verification failures rather than broadly expanding cultural details; the same evidence space also represents avoid constraints for cross-cultural mixing, hallucinated symbols, and stereotyped or over-symbolized cues.

3 Method

3.1 Problem Formulation

Given a cultural activity prompt p, our goal is to generate an image I that faithfully depicts the activity. Unlike culture noun generation, this requires joint control over visible objects, participant roles, actions, interactions, scene context, and visual elements to avoid. CARE formulates the task as evidence-grounded generation under visual constraints: it derives source-grounded evidence and checkable requirements, then uses them for generation, verification, and repair. Each prompt is decomposed into retrieval queries Q, used to retrieve evidence records R, and converted into an Evidence Card E and checklist C:

$$\begin{aligned} Q &= f_{\mathrm{q}}(p), \qquad R = f_{\mathrm{ret}}(Q), \\ E &= f_{\mathrm{ev}}(p, R), \qquad C = f_{\mathrm{ck}}(E). \end{aligned} \tag{1}$$

Here, f_{q} denotes query decomposition, f_{ret} retrieval, and f_{ev} evidence filtering and organization. The Evidence Card stores visualizable, checkable, source-traceable evidence, and the checklist turns it into local VLM-verifiable questions.

3.2 Overall Framework

Figure 2 gives an overview of CARE. The framework is organized as a closed-loop pipeline that keeps the same evidence and checklist representation throughout generation, evaluation, and repair. Given an input cultural activity prompt p, CARE first constructs an evidence space that specifies what should be visually present and what should be avoided. This evidence is not used as a loose textual addition to the prompt; instead, it is converted into two linked artifacts: an initial generation prompt p_0 and a checklist C_0 whose items can be verified from the generated image. The pipeline contains three stages: Stage 1 retrieves and filters cultural evidence, organizes it into an Evidence Card, and derives p_0 and C_0; Stage 2 generates an initial image I_0 from p_0 and verifies it with VLM checklist questions; Stage 3 uses failed checklist items and linked Evidence Card atoms to construct targeted repair prompts, generate multiple candidates, recheck relevant constraints, and select the candidate that best repairs failures while avoiding regressions.

This design differs from retrieval-augmented prompting in that retrieved evidence is transformed into a stable verification space rather than simply appended to the generation prompt. It also differs from standard best-of-K selection, which usually ranks candidates by a global preference score: CARE explicitly tracks which constraints have already passed and prioritizes candidates that fix remaining failures without discarding previously satisfied evidence. Ours(1) uses only the initial generation–verification pass without repair, whereas Ours(3) allows up to three evidence-guided repair rounds. The Evidence Card and checklist are constructed once and remain fixed across all rounds.

3.3 Stage 1: Evidence Card and Checklist Construction

Stage 1 constructs a stable evidence space. We use a task-driven visual decomposition rather than a universal cultural taxonomy: objects capture artifacts, actions capture behaviors and interactions, scenes capture context, and avoid captures cultural mixing, hallucinations, and stereotypes. This design follows activity-centered cultural evaluation and fine-grained visual error localization (Malakouti et al., 2026; Nayak et al., 2025; Hu et al., 2023).

An LLM decomposes p into queries for these four slots. CARE then performs slot-wise web retrieval, deduplication, and relevance ranking; during evaluation, retrieval does not access the sample-specific source/evidence spans used to construct CARE-1360. The filtered records become an Evidence Card whose atoms have identifiers, source links, observable visual content, and corresponding VLM-checkable questions. Atoms are marked `checkable=true` only when they can be judged visually from a single image without inferring identity, dialogue, intent, language, or external knowledge. `checkable=false` atoms are retained only

for provenance and do not enter the checklist, verification, scoring, stopping, or repair. The card yields p_0 and fixed C_0, serving as the shared representation for prompt optimization, verification, and repair; Appendix H gives examples and implementation details.

We construct p_0 from the original activity and the visual forms of object, action, and scene atoms, with negative constraints from avoid atoms. C_0 includes only `checkable=true` atoms: positive items expect `yes`, while avoid items expect `no`. It remains fixed across repair rounds.

3.4 Stage 2: Checklist-Based Image Verification

Stage 2 generates an image from p_0 and verifies it with a VLM. Rather than using holistic 1–5 cultural-correctness scores inside the loop, CARE applies constrained yes/no checklist verification. For item c_j with expected answer $\hat{a}_j$:

$$r_j = \mathbb{I}[a_j = \hat{a}_j], \tag{2}$$

where $a_j \in \{\texttt{yes}, \texttt{no}\}$.

Following fine-grained VQA evaluation (Hu et al., 2023), binary judgments provide actionable control signals by mapping each failure to a specific repair target. Unlike holistic Likert scores, they avoid conflating partial satisfaction, severity, and preference. We retain 1–5 ratings for final human evaluation, where graded judgments are more appropriate.

Let C_{req} comprise all positive checkable object, action, and scene items; $C_{\text{act}} \subseteq C_{\text{req}}$ contains the action items, and C_{avoid} the checkable avoid items. Required actions are visually judgeable, supported by the prompt or retrieved evidence, and expect `yes`.

The report builder aggregates item results into coverage, avoid status, action status, failed sources, Evidence Recall notes, and prioritized failures. Required coverage is

$$\text{Cov}(I, C) = \frac{\sum_{j \in C_{\text{req}}} r_j}{|C_{\text{req}}|}. \tag{3}$$

Stopping requires sufficient coverage, all avoid items to pass, and all required action items to pass, making verification a structured repair target.

3.5 Stage 3: Evidence-Guided Iterative Repair

If the image fails verification, Stage 3 repairs it while keeping C_0 fixed, preventing the system from redefining cultural correctness across iterations. CARE maintains a global state $s_j^{(t)}$ for each checklist item. At iteration t, only an active subset $C_{\text{active}}^{(t)}$ is rechecked:

$$s_j^{(t)} = \begin{cases} r_j^{(t)}, & c_j \in C_{\text{active}}^{(t)}, \\ s_j^{(t-1)}, & c_j \notin C_{\text{active}}^{(t)}. \end{cases} \tag{4}$$

The active checklist contains failed or unknown items, all avoid and required action items, and a few passed items from failing slots for preservation. The LLM receives these failures, linked Evidence Card atoms, preservation notes, avoid constraints, and a bounded subset of filtered passages, then generates multiple repair candidates. Candidate selection uses a target-aware best-of-K score that rewards required coverage (C_k), repaired failures (F_k), core-action success (A_k), and preservation (P_k), while penalizing avoid violations (V_k) and regressions (R_k):

$$S_k = \alpha C_k + \beta F_k + \gamma A_k + \delta P_k - \eta V_k - \mu R_k. \tag{5}$$

The best candidate is returned if it meets the stopping condition; otherwise, the state is updated until the iteration budget is exhausted or no improvement is observed.

3.6 Cultural Activity Prompt Dataset

For evaluation, we construct CARE-1360 from Cultural Atlas (Mosaica, 2024), focusing on visually depictable activities rather than isolated cultural nouns. It covers eight countries—China, Canada, Brazil, Japan, Iran, South Africa, Poland, and Germany—and five categories: dates of significance, etiquette, family life, religion, and greetings. We extract country-category activities, filter entries that are abstract or hard to visualize, generate source-grounded prompts, and balance country coverage through semantic-preserving expansion, near-duplicate filtering, and quality checks. CARE-1360 contains 1,360 traceable prompts, with 170 per country; category counts remain non-uniform to preserve source availability and visual-depictability filtering (Table 1). Audits show limited augmentation (15 prompts, 1.10%), unique source record identifiers, no exact or normalized duplicate prompts, and 909 unique activity names; full construction details are provided in Appendix A, with summary statistics in Section A.8.

Country	Dates	Etiq.	Family	Greet.	Relig.	Total
Brazil	21	75	30	24	20	170
Canada	61	69	23	8	9	170
China	30	66	36	13	25	170
Germany	64	73	15	8	10	170
Iran	28	72	34	15	21	170
Japan	18	84	21	21	26	170
Poland	33	78	24	28	7	170
South Africa	45	85	0	23	17	170
Total	300	602	183	140	135	1360

Table 1: Distribution of CARE-1360 by country and activity category. Each country contributes 170 prompts. Category counts reflect Cultural Atlas source availability and visual-depictability filtering; South Africa has no Family records because the corresponding Cultural Atlas family page has no available source records.

4 Experiments

4.1 Experimental Setup

We conduct experiments on CARE-1360, covering eight countries: Brazil, Canada, China, Germany, Iran, Japan, Poland, and South Africa. Each country contains 170 cultural activity prompts, yielding 1,360 samples in total. Our main experiments use ZImage (Team et al., 2026) as the unified generation backend to generate 1024×1024 images, ensuring that comparisons primarily reflect differences in prompt construction and optimization strategies. To examine whether CARE is specific to this backend, we further conduct a robustness evaluation on Stable Diffusion XL (SDXL) Base 1.0 (Podell et al., 2024), a widely used latent diffusion model for high-resolution text-to-image synthesis. All compared methods use the same prompts, image resolution, and human evaluation protocol across backends. Online evidence retrieval uses Wikipedia and DuckDuckGo as search sources, and no method is given direct access to the sample-specific source passages or evidence spans used during dataset construction. Unless otherwise specified, query decomposition, LLM synthesis, and refiner components are implemented with DeepSeek-V3.2 (671B) (DeepSeek-AI et al., 2025), checklist-based image verification uses Qwen3-VL-32B-Instruct (Bai et al., 2025a; Yang et al., 2025; Bai et al., 2025b; Wang et al., 2024; Bai et al., 2023), and embedding-based deduplication and similarity filtering use Qwen/Qwen3-Embedding-8B (Zhang et al., 2025b).

Stopping and repair parameters. CARE enters Phase 3 only when the initial image fails the shared stopping rule. A candidate meets this rule when required coverage is at least 0.75, all avoid items pass, and all required action items pass. Otherwise, repair stops at the iteration budget or after two consecutive rounds without improvement, defined as no increase in coverage and no newly satisfied avoid or core-action status. Each repair round generates $K = 3$ candidates. Ours(1) uses only the initial pass, while Ours(3) allows up to three repair rounds.

4.2 Human Evaluation Protocol

We collect human annotations through a crowdsourcing platform. The evaluation covers the 1,360 cultural activity prompts in CARE-1360 under five generation methods, yielding 6,800 image–prompt pairs. For each of the eight countries, we recruit five annotators who are familiar with the corresponding cultural context, resulting in 40 annotators in total; each annotator only evaluates samples from the country they are familiar with, so each image–prompt pair receives five independent ratings, which are averaged as the final score.

Annotators evaluate each image with respect to the original cultural activity prompt along six dimensions: Overall Activity Faithfulness, Role Correctness, Action Correctness, Hallucination, Scene Layout & Temporal Setting, and Stereotype Intensity. All dimensions use a 1–5 scale, where higher scores indicate better quality; higher Hallucination scores mean fewer incorrect or cross-cultural elements, and higher Stereotype Intensity scores mean more natural and less overly symbolic depictions. We further compute within-country inter-annotator agreement among the five annotators assigned to each country; Pearson correlations range from 0.790 to 0.920, Spearman correlations from 0.754 to 0.912, and Krippendorff's alpha from 0.789 to 0.910. Details of annotator screening, interface design, randomization, quality control, and agreement statistics are provided in Appendix I.

4.3 Compared Methods

We compare the following five methods:

- **Base**: Directly uses the original cultural activity prompt to generate an image, without additional retrieval or prompt optimization.
- **C-TRIP(1)**: Uses the unchanged official C-TRIP codebase with up to one round of

prompt refinement. Because C-TRIP was designed for culture-specific nouns, we use a lightweight data-normalization adapter that converts cultural activity prompts into C-TRIP-compatible inputs, enabling its retrieval-based cultural-background and visual-detail prompting to run on our activity prompts.

- **C-TRIP(3)**: Uses the same unchanged official C-TRIP codebase and adapter with up to three rounds of prompt refinement, allowing iterative prompt rewriting based on retrieved cultural knowledge.
- **Ours(1)**: Uses one round of our evidence- and checklist-guided generation process.
- **Ours(3)**: Uses CARE with up to three rounds of repair and candidate selection.

The number in parentheses denotes the maximum number of allowed iteration rounds; for example, (1) means only the initial round is performed, while (3) allows up to three rounds of refinement or repair. For fairness, the C-TRIP baselines use the same activity prompts and image-generation backend as the other methods, while keeping the official C-TRIP code unchanged. The adapter only reformats activity prompts into C-TRIP-compatible inputs; it does not add task-specific evidence or alter the baseline algorithm. Neither C-TRIP nor CARE receives the original Cultural Atlas evidence span as an input field; differences therefore reflect the retrieval and prompt-construction procedures available to each method rather than privileged reuse of dataset construction records.

We set the default candidate-search budget to $K = 3$ because the validation ablation in Figure 3 shows that most gains appear by the third candidate. Avg6 increases from 4.319 at $K = 1$ to 4.389 at $K = 3$, then saturates at 4.389–4.390 for $K = 4$–5, suggesting that larger candidate pools add little benefit relative to their extra generation cost. The candidate-selection weights in Eq. (5) are fixed heuristic parameters rather than learned values, and are set to match the hard-coded scoring rule used in our implementation. Specifically, C_k is the required-checklist coverage in $[0, 1]$; F_k counts how many items that failed in the previous state are repaired by the current candidate; $A_k \in \{0, 1\}$ indicates whether all core-action items pass; P_k counts Evidence Recall preservation items that still pass, where these items are previously satisfied

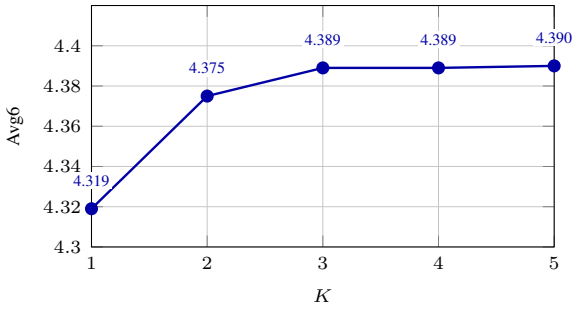


Figure 3: Avg6 under different numbers of candidate images on the 480-prompt ZImage subset.

evidence-grounded checklist items retained in the repair prompt so that the model does not discard already recalled evidence while fixing new failures; $V_k \in \{0, 1\}$ indicates whether the candidate contains any avoid violation; and R_k counts regressions, i.e., items that passed in the previous state but fail in the current candidate. The implemented selection score is therefore

$$S_k = 10C_k + 2.0F_k + 1.0A_k + 0.5P_k - 5.0V_k - 0.5R_k,$$

which corresponds to

$$(\alpha, \beta, \gamma, \delta, \eta, \mu) = (10, 2.0, 1.0, 0.5, 5.0, 0.5)$$

in Eq. (5). Thus, coverage receives the largest positive weight, repairing an originally failed item is weighted more strongly than preserving an already passed item, a single avoid violation receives a large penalty, and regressions are penalized individually. A sensitivity analysis on 200 prompts, evaluated by GPT-5.5 with the backend, $K = 3$, and maximum repair rounds fixed, shows little difference between the default and repair-heavy settings (4.1233 vs. 4.1250 Avg6), while the coverage-only setting performs worst (4.0717). This suggests that CARE is not brittle to the default weights and that target-aware signals beyond coverage are useful; full results are provided in Appendix B.

5 Results and Analysis

5.1 Main Results

Table 2 reports human evaluation on ZImage and SDXL Base 1.0 (Podell et al., 2024). Enhanced methods generally improve activity faithfulness, role correctness, action correctness, and scene/temporal setting over Base, indicating that

Backend	Method	Faithfulness ↑	Role ↑	Action ↑	Hallucination ↑	Scene ↑	Stereotype ↑	Avg6 ↑
ZImage	Base	3.929	4.329	3.940	4.281	4.079	4.303	4.144
	C-TRIP(1)	4.088	4.415	4.057	4.286	4.221	4.017	4.181
	C-TRIP(3)	4.079	4.396	4.036	4.254	4.218	3.931	4.152
	Ours(1)	4.215	4.550	4.250	4.429	4.352	4.248	4.341
	Ours(3)	**4.260**	**4.592**	**4.307**	**4.479**	**4.377**	**4.308**	**4.387**
SDXL	Base	1.366	1.512	1.248	1.449	1.369	1.946	1.482
	C-TRIP(1)	1.678	1.908	1.478	1.900	1.719	2.511	1.866
	C-TRIP(3)	2.118	2.281	1.863	2.461	**2.301**	3.209	2.372
	Ours(1)	2.086	2.417	1.825	2.461	2.160	3.443	2.399
	Ours(3)	**2.125**	**2.419**	**1.867**	**2.476**	2.225	**3.532**	**2.441**

Table 2: Human evaluation results using ZImage and SDXL Base 1.0 as image-generation backends. Higher scores are better for all dimensions. ZImage results are computed over 1,360 prompts; full bootstrap 95% confidence intervals and paired comparisons for ZImage are provided in Appendix J.

activity-oriented prompt enrichment helps generated images express cultural activity contexts. Figure 1 shows one example; more are in Appendix G.

On ZImage, Ours(3) is highest on all six dimensions and Avg6, improving over Base, C-TRIP(1), C-TRIP(3), and Ours(1) by +0.244, +0.206, +0.235, and +0.046 Avg6 points, respectively; paired differences are significant (Table 17). The gains are modest on the 1–5 scale, so we interpret CARE as a structured control mechanism that yields consistent incremental improvements. SDXL shows the same overall trend, suggesting transfer across the tested generation backends.

VIEScore (Ku et al., 2024) is used only as a sanity check for generic prompt–image alignment and perceptual quality, since it does not directly measure cultural activity grounding, stereotype control, or role–action–scene correctness; details are in Appendix K.

Table 3 shows that CARE maintains comparable generic VIEScore scores to baselines. Ours(1) is marginally higher than Ours(3) on SC, PQ, and Overall, while Ours(3) performs better in human evaluation. This contrast suggests that the additional human-evaluation gains come from cultural-activity criteria rather than generic T2I quality.

5.2 Ablation and Analysis

Component ablation. Table 4 disentangles CARE components on a balanced 480-prompt ZImage subset. CARE (evidence only) raises Avg6 from 4.221 for Base and 4.251 for C-TRIP(1) to 4.396, showing that activity-oriented evidence provides a useful generation signal. CARE (full) further improves Avg6 to 4.454. Because the verifier still triggers repair for 36.7% of samples, especially on core actions and avoid constraints, the repair loop adds value by localizing unresolved visual requirements and using them for targeted prompt revision and candidate selection. Detailed trigger rates, stopping reasons, and image-generation costs are reported in Appendix C.

Candidate-search capacity. Figure 3 shows that Avg6 rises from 4.319 at $K = 1$ to 4.389 at $K = 3$, then saturates at 4.389–4.390 for $K = 4$–5. Thus, CARE's benefit is not simply unbounded sampling; most gains come from evidence-grounded failure localization and target-aware selection.

Overall, CARE improves through evidence-grounded prompting, checklist-based diagnosis, and targeted repair: evidence alone helps, while verification and repair convert unresolved constraints into additional but modest gains.

5.3 Country-Level Observations

At the country level, Ours(3) achieves the highest Avg6 score in most countries except China, where Ours(1) is slightly higher. Table 14 shows positive gains over Base for all eight countries, with larger gains for Iran and Poland and smaller gains for Germany, Japan, Brazil, and South Africa.

Canada and Iran obtain relatively lower absolute scores, possibly because many samples involve abstract social norms, everyday etiquette, or weak visual cues. Countries such as China, Japan, and Brazil include more festivals, family interactions, or concrete etiquette practices, which are easier to benefit from structured evidence. Full per-country variance and standard-deviation statistics are provided in Table 16. Overall, activities with clear ritual actions, visible objects, and scene structures are easier than those involving abstract social norms, identity relations, or weak visual cues.

Method	Base	C-TRIP(1)	C-TRIP(3)	Ours(1)	Ours(3)
Semantic Consistency (SC) ↑	0.782	0.791	0.791	**0.803**	0.798
Perceptual Quality (PQ) ↑	8.372	8.311	8.272	**8.417**	8.394
VIEScore Overall ↑	0.806	0.807	0.805	**0.819**	0.815

Table 3: Generic T2I quality sanity check with VIEScore using GPT-5 (Singh et al., 2026) as the evaluation model. SC and Overall are normalized to 0–1, while PQ is reported on the original 0–10 scale. Higher scores indicate better generic prompt–image alignment or perceptual quality, not cultural correctness.

Method	N	Faith.	Role	Action	Halluc.	Scene	Stereo.	Avg6
Base	480	4.023	4.367	4.004	4.381	4.148	**4.404**	4.221
C-TRIP(1)	480	4.167	4.446	4.129	4.388	4.302	4.075	4.251
C-TRIP(3)	480	4.098	4.392	4.040	4.273	4.238	3.942	4.164
CARE (evidence only)	480	4.283	4.583	4.319	4.510	4.408	4.269	4.396
CARE (full)	480	**4.358**	**4.635**	**4.400**	**4.540**	**4.456**	4.335	**4.454**

Table 4: Component ablation on a balanced 480-prompt subset using ZImage as the image-generation backend. The subset contains 60 prompts per country across the eight countries. All methods are evaluated with the same six-dimensional human scoring protocol as the main experiment. CARE (evidence only) uses evidence-grounded prompting without checklist-guided repair, while CARE (full) uses the complete verification and repair loop.

6 Conclusion

This paper presents CARE, an evidence-driven visual checklist framework for cultural activity image generation. Unlike methods that focus on cultural nouns or static cultural objects, we emphasize that cultural activities require joint modeling of participants, actions, interactions, scenes, and avoidance constraints. CARE converts activity prompts into source-traceable cultural evidence and fine-grained visual checklists, and uses VLM-identified failed checks to guide targeted prompt repair. Experiments suggest that CARE provides consistent but moderate gains over direct generation and Culture-TRIP-style refinement on both ZImage and SDXL backends, especially in activity faithfulness, action correctness, scene consistency, hallucination control, and stereotype control. Future work can extend the framework to more countries and more complex cross-cultural scenarios, and explore more reliable automated evidence retrieval and verification mechanisms.

Limitations

This work has several limitations. CARE-1360 covers eight countries and five broad categories of cultural activities, which enables controlled comparison but cannot capture the full regional, ethnic, religious, linguistic, socioeconomic, and diasporic diversity within each country; therefore, the evidence cards and generated images should not be interpreted as canonical representations of national cultures. Our evaluation also relies on both human annotation and VLM-based checking, which are inherently subjective and may reflect disagreement, recognition errors, or cultural biases despite the use of unified rubrics and country-specific annotators. Finally, experiments are limited to ZImage and SDXL Base 1.0, so future work should test newer proprietary models, video generation systems, and interactive editing pipelines, while expanding cultural coverage and incorporating community review of evidence cards.

Ethical Considerations

We have not identified any ethical concerns directly related to this study.

Acknowledgments

This work was supported by Tianshan Talent Training Program (2023TSYCCX0041); the Key Research and Development Program of Xinjiang Uyghur Autonomous Region (Grant No. 2024B03026, Grant No. 2023B03024); the Natural Science Foundation of Xinjiang Uyghur Au-

tonomous Region, (Grant No. 2022D01D81); National Key Research and Development Program of China (Grant No. 2025YFC3309300); the Outstanding Member Program of the Youth Innovation Promotion Association of Chinese Academy of Sciences (Grant No. Y2023118); Key Science and Technology Innovation Program of the Chinese Academy of Sciences (No. KGFZD-145-26-44); the Youth Talents Support Project of Xinjiang Uyghur Autonomous Region (2023TSYCQNTJ0037); the Key Research and Development Program of Xinjiang Uyghur Autonomous Region, (Grant No. 2025B03042); Guidance on Local Science and Technology Development Fund of Xinjiang Uygur Autonomous Region (Grant No. ZYYD2026JD03).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, volume 2024, pages 9112–9141.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL technical report. *Preprint*, arXiv:2511.21631.

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

Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5136–5147.

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. BeautifulPrompt: Towards automatic prompt engineering for text-to-image synthesis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1–11, Singapore. Association for Computational Linguistics.

Jaemin Cho, Yushi Hu, Jason Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*, volume 2024, pages 15625–15645.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. DALL-Eval: Probing the reasoning skills and social biases of Text-to-Image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3043–3054.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 78723–78747. Curran Associates, Inc.

Suchae Jeong, Inseong Choi, Youngsik Yun, and Jihie Kim. 2025. Culture-TRIP: Culturally-aware text-to-image generation with iterative prompt refinement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9543–9573.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. In *Advances in Neural Information Processing Systems*, volume 37, pages 13716–13747. Curran Associates, Inc.

Mohammed Abdul Hafeez Khan, Yash Jain, Siddhartha Bhattacharyya, and Vibhav Vineet. 2025. Test-time prompt refinement for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 6565–6575.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance.

In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2024. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12268–12290, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. 2024. GenAI-Bench: Evaluating and improving compositional text-to-visual generation. *Preprint*, arXiv:2406.13743.

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023. On the cultural gap in text-to-image generation. *Preprint*, arXiv:2307.02971.

Zhixuan Liu, Peter Schaldenbrand, Beverley-Claire Okogwu, Wenxuan Peng, Youngsik Yun, Andrew Hundt, Jihie Kim, and Jean Oh. 2024. SCoFT: Self-contrastive fine-tuning for equitable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10822–10832.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 56338–56351. Curran Associates, Inc.

Sina Malakouti, Boqing Gong, and Adriana Kovashka. 2026. Culture in action: Evaluating text-to-image models through social activities. In *The Fourteenth International Conference on Learning Representations*.

Mosaica. 2024. The cultural atlas. `https://culturalatlas.sbs.com.au/`. Accessed May 1, 2026.

Shravan Nayak, Mehar Bhatia, Xiaofeng Zhang, Verena Rieser, Lisa Anne Hendricks, Sjoerd van Steenkiste, Yash Goyal, Karolina Stanczak, and Aishwarya Agrawal. 2025. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20918–20953, Suzhou, China. Association for Computational Linguistics.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pages 1862–1874.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Aniket Rege, Zinnia Nie, Mahesh Ramesh, Unmesh Raskar, Zhuoran Yu, Aditya Kusupati, Yong Jae Lee, and Ramya Korlakai Vinayak. 2025. CuRe: Cultural gaps in the long tail of text-to-image systems. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15680–15691.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22227–22243, Suzhou, China. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. OpenAI GPT-5 system card. *Preprint*, arXiv:2601.03267.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, and 5 others. 2026. Z-Image: An efficient image generation

foundation model with single-stream diffusion transformer. *Preprint*, arXiv:2511.22699.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shih-Ying Yeh, Yi Li, SangHyun Park, Giyeong Oh, Xuehai Wang, Min Song, Youngjae Yu, and Shang-Hong Lai. 2026. TIPO: Text to image with text presampling for prompt optimization. In *International Conference on Learning Representations*, volume 2026, pages 131998–132048.

Jian Zhang, Junyi Guo, Junyi Yuan, Huanda Lu, Yanlin Zhou, Fangyu Wu, Qiufeng Wang, and Dongming Lu. 2025a. LLM-driven completeness and consistency evaluation for cultural heritage data augmentation in cross-modal retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19407–19417, Suzhou, China. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Dataset Construction Details

A.1 Source Collection

We construct our cultural activity prompts from Cultural Atlas country–category pages. For each country, we consider five broad categories: dates of significance, etiquette, family life, religion, and greetings. Unlike object-centric cultural prompt sets, our goal is to collect visually depictable cultural activities: events, interactions, rituals, greetings, household practices, religious participation, and other scenes in which people perform culturally meaningful actions.

For reproducibility, we store all extracted records with their source URL, section title, source text, and evidence span. Each candidate activity is therefore traceable to a source passage rather than being generated from an ungrounded cultural label. These source passages and evidence spans are not exposed to the generation methods during evaluation; they are used for dataset construction, traceability, and auditing only.

A.2 Activity Candidate Extraction

From each Cultural Atlas page, we extract candidate activity records. A candidate is retained when it describes or implies a visually observable activity, rather than an abstract value, isolated cultural noun, or static object. For example, we retain entries involving greeting behavior, food sharing, festival participation, visiting etiquette, religious observance, and family interaction. We exclude entries that cannot be meaningfully depicted in an image without adding unsupported assumptions.

Each retained record contains the target country, content category, normalized activity name, source text, evidence text, source URL, and a stable record identifier. The activity name is used as the main cultural anchor for later prompt generation and evidence retrieval.

A.3 Evidence-Grounded Prompt Generation

For each retained activity, we generate an initial text-to-image prompt grounded in the extracted source evidence. The prompt is required to describe a concrete visual scene, including participants, actions, setting, and culturally relevant objects when these are supported by the evidence. The generator is instructed to avoid adding unsupported cultural symbols or generic stereotypes.

In addition to the prompt itself, we store retrieval-oriented fields for each record, including primary

search terms, contextual search terms, and visual-detail search terms. These fields are later used by the evidence retrieval and checklist construction modules. The original source text and evidence span are kept with every prompt as dataset metadata for traceability, but are not provided as direct inputs to the compared generation methods.

A.4 Filtering and Quality Control

We apply several quality-control steps after prompt generation. First, prompts with malformed outputs or missing required fields are removed. Second, prompts are audited for cultural anchoring, visual observability, and consistency with the source evidence. Prompts that are too abstract, too generic, weakly tied to the target country, or overly dependent on unsupported cultural assumptions are flagged for revision or removal.

For flagged prompts that can be repaired, we rewrite them while preserving the original activity and evidence anchor. The rewrite step is constrained to improve visual concreteness and cultural grounding, rather than introducing new cultural facts. For example, a vague prompt may be rewritten to specify the participants, action, and setting, while keeping the same source-supported activity.

A.5 Semantic-Preserving Expansion

Because Cultural Atlas pages differ in coverage across countries and categories, some countries provide fewer directly usable activity prompts than others. To reduce imbalance, we perform semantic-preserving expansion for underrepresented portions of the data. These variants keep the same source activity and evidence anchor, but vary the visual realization of the scene.

The allowed variation includes changes in viewpoint, participant arrangement, action detail, object emphasis, or scene composition. For example, a family festival activity may be represented through different family members participating in the same source-supported practice, or through a closer view of the relevant cultural action. The expansion is not allowed to change the cultural activity itself or introduce unsupported traditions. Each variant retains a link to the parent source record.

A.6 Deduplication and Final Selection

After expansion, we remove near-duplicate prompts to improve diversity. We use embedding similarity to identify prompts that are semantically too close, and retain prompts that provide distinct visual realizations of the same or related cultural activity. The final selection balances country coverage while preserving activity diversity across the five content categories.

The resulting dataset consists of traceable cultural activity prompts. Each prompt includes the target country, category, activity name, source evidence, source URL, and a stable identifier. This design allows the dataset to support both image generation and evidence-grounded evaluation: generated images can be assessed not only against the prompt, but also against the cultural evidence from which the prompt was derived.

Metric	Value
Rows	1,360
Countries	8
Categories	5
Unique activity names	909
Unique source texts	585
Unique evidence texts	911
Unique source URLs	39
Unique source record IDs	1,360
Section titles	51

Table 5: Basic metadata statistics for CARE-1360. Source URLs correspond to broader Cultural Atlas pages, while source record IDs are prompt-level provenance identifiers.

A.7 Final Category Distribution

Table 1 in the main text reports the final country–category counts. The selection keeps country totals balanced at 170 prompts per country, but does not force the five activity categories to be uniform within each country. Instead, the final distribution reflects two constraints: the amount of usable Cultural Atlas material available for each country–category page and whether extracted activities can be converted into concrete visual scenes without adding unsupported cultural assumptions. Consequently, Etiquette contributes the largest number of prompts overall, while Dates of Significance is relatively prominent for Canada and Germany. South Africa has no Family prompts because the corresponding Cultural Atlas family page has no available source records.

A.8 Dataset Statistics

Table 5 summarizes the basic metadata coverage of CARE-1360. The dataset contains 1,360 rows across eight countries and five categories, with 909 unique activity names and 911 unique evidence

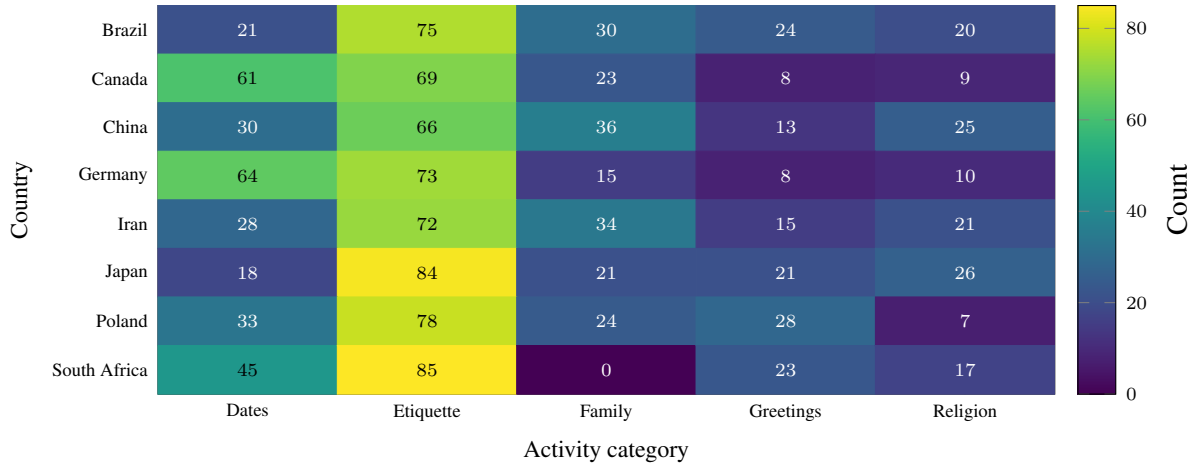


Figure 4: Heatmap of the final CARE-1360 country–category distribution. Darker cells indicate more prompts in that category. The zero Family count for South Africa reflects that the corresponding Cultural Atlas family page had no available source records, rather than post-hoc removal of otherwise usable records.

texts. Although the prompts are derived from 39 Cultural Atlas source URLs, the final rows use 1,360 unique source record identifiers; this means that the final prompt set does not reuse the same source record as a direct duplicated parent. Source URLs are therefore best interpreted as country–category source pages, while source records and evidence texts capture finer-grained activity-level provenance.

Table 6 reports overlap statistics under different grouping keys. The source-record grouping has singleton rate 1.000 and maximum cluster size 1, confirming that the final dataset does not contain multiple prompts with the same source record identifier. The source-text and activity-name groupings show natural overlap because a source passage may support multiple distinct visual activities and the same activity name may occur in different contexts; however, the maximum source-text and activity-name cluster sizes are both 7, rather than forming large template-like clusters.

Table 7 reports word-length statistics for key textual fields. Prompts are relatively concise, with a mean length of 18.66 words, while source texts are longer and more variable. This reflects the construction process: prompts are compact visual scene descriptions, whereas source texts preserve broader contextual passages for traceability and auditing.

Table 8 reports exact and normalized duplicate statistics. The prompt fields contain no exact or normalized duplicates. Repetition at the activity-name, source-text, and evidence-text levels reflects the structure of the source material: the same activity label or source passage can support multiple distinct prompts, but the final prompt strings themselves are unique.

Cluster key	Groups	Mean	Median	P90	Max
Source record ID	1,360	1.000	1	1	1
Source text	585	2.325	2	4	7
Source URL	39	34.872	25	73.4	85
Activity name	909	1.496	1	2	7

Table 6: Source-overlap and cluster-size statistics for CARE-1360. P90 denotes the 90th percentile of cluster size. Source URL clusters are larger because each URL corresponds to a broader country–category page; source-record clusters are singleton by construction. Singleton rates are 1.000 for source record ID, 0.352 for source text, 0.000 for source URL, and 0.569 for activity name.

Field	Mean words	Median	P90	Max
Prompt	18.66	18	24	49
Source text	42.65	28	105	190
Evidence text	8.93	8	15	30
Activity name	3.93	4	6	10

Table 7: Word-length statistics for CARE-1360 textual fields. P90 denotes the 90th percentile, meaning that 90% of entries in the corresponding field are no longer than this value.

Table 13 reports two additional dataset attributes. Most prompts describe social practices rather than named events, which supports the focus on culturally grounded activities beyond festival names or static cultural nouns. The confidence field records the construction-time confidence assigned during dataset preparation; because it is metadata rather than an evaluation outcome, we use it descriptively

Field	Unique	Duplicate rows	Dup. rate	Max dup.
Prompt	1,360	0	0.000	1
Final prompt	1,360	0	0.000	1
Normalized prompt	1,360	0	0.000	1
Activity name	909	451	0.332	7
Source text	585	775	0.570	7
Evidence text	911	449	0.330	6

Table 8: Duplicate statistics for CARE-1360. Prompt-level duplicates are absent under exact and normalized matching; source-level repetitions reflect shared source passages or activity labels rather than duplicate prompt strings.

Setting	$(\alpha, \beta, \gamma, \delta, \eta, \mu)$	Avg6 ↑
Default	$(10, 2, 1, 0.5, 5, 0.5)$	4.1233
Repair-heavy	$(8, 4, 1, 0.5, 5, 0.5)$	**4.1250**
Balanced	$(5, 1, 1, 1, 5, 1)$	4.0992
Avoid-heavy	$(10, 2, 1, 0.5, 10, 0.5)$	4.0975
Coverage-only	$(10, 0, 0, 0, 0, 0)$	4.0717

Table 9: Sensitivity of the Eq. (5) selection weights on 200 prompts, evaluated by GPT-5.5. The backend, $K = 3$, and maximum repair rounds $= 3$ are fixed across settings.

rather than as a filtering criterion in the experiments.

B Selection-Weight Sensitivity

We evaluate the sensitivity of the candidate-selection score in Eq. (5) on 200 prompts sampled from the balanced 480-prompt ZImage ablation subset. We fix the image-generation backend, $K = 3$, and maximum repair rounds to three, and vary only $(\alpha, \beta, \gamma, \delta, \eta, \mu)$, which weight required coverage, repaired failures, core-action success, preservation, avoid violations, and regressions, respectively. All settings are evaluated by GPT-5.5 under the same six-dimensional rubric.

Table 9 shows that the repair-heavy setting differs from the default by only $+0.0017$ Avg6, while the balanced and avoid-heavy settings decrease modestly. Coverage-only performs worst, indicating that repaired failures, core-action success, preservation, avoid violations, and regressions provide useful selection signals beyond aggregate coverage. Overall, performance is stable across several reasonable weight configurations rather than depending on a single manual setting.

C Adaptive Repair Cost Analysis

CARE applies repair conditionally rather than assigning the maximum repair budget to every prompt. On the balanced 480-prompt subset in Table 4, 304 prompts (63.3%) satisfy the stopping condition after the initial generation–verification pass, while 176 prompts (36.7%) trigger repair. Consequently, the gains from full CARE over evidence-only prompting are averaged over both repaired samples and the larger group of early-pass samples. Nevertheless, full CARE improves Faithfulness by $+0.075$, Role by $+0.052$, Action by $+0.081$, Scene by $+0.048$, and Stereotype by $+0.067$ over evidence-only prompting.

The realized image-generation cost is substantially below the worst-case budget. The subset requires 480 initial images and 750 repair candidates, for 1,230 images in total, or 2.56 images per prompt on average. This is lower than applying the maximum budget of $1+3\times3 = 10$ images to every prompt. Among the 176 repaired samples, 79.5% stop after meeting the coverage condition, 9.1% stop after two consecutive rounds without improvement, and 11.4% reach the maximum repair-round budget. These results show that additional generation is concentrated on checklist-identified failures rather than uniformly applied to all samples.

D Additional Diagnostic Baselines

To distinguish CARE from retrieval-free prompt expansion and generic best-of-K selection, we introduce two additional diagnostic baselines. **No-Retrieval LLM Expansion** does not use web retrieval, Evidence Cards, checklists, or repair. It asks the same LLM to rewrite the original activity prompt with concrete visual details about participants, actions, objects, and scene context, and generates one image from the expanded prompt. **Generic VLM Best-of-K** uses CARE's evidence-grounded initial prompt and generates $K = 3$ candidates, but removes the atomic checklist and evidence-linked repair process. Instead, it selects a candidate using a generic holistic judgment from the same VLM.

We evaluate these baselines on 200 prompts sampled from the balanced 480-prompt ZImage subset used in Table 4. Because the rebuttal period did not permit culturally familiar human evaluation at the scale and rigor of the main experiment, we use GPT-5.5 as a temporary automatic judge under the same six-dimensional rubric. These results are reported separately from the main human-evaluation results and should be interpreted as diagnostic evidence rather than as a replacement for country-familiar

Method	Retrieval	Checklist	Repair	K	Avg6 ↑
No-Retrieval LLM Expansion	No	No	No	1	4.241
Generic VLM Best-of-K	Yes	No	No	3	4.173
Full CARE	Yes	Yes	Yes	3	**4.447**

Table 10: Diagnostic evaluation of retrieval-free prompt expansion and generic VLM best-of-K selection on 200 prompts. Avg6 is assigned by GPT-5.5 under the six-dimensional evaluation rubric. For Full CARE, $K = 3$ denotes the number of candidates generated in each triggered repair round; repair is applied adaptively as described in Appendix C.

human evaluation.

As shown in Table 10, Full CARE exceeds Generic VLM Best-of-K and No-Retrieval LLM Expansion by $+0.274$ and $+0.206$ Avg6, respectively. This diagnostic comparison suggests that CARE's gains cannot be explained solely by retrieval-free activity-prompt expansion, multi-candidate sampling, or generic VLM reranking. Instead, the combination of evidence grounding, checklist-based failure localization, and targeted repair provides additional value. Because these results rely on automatic evaluation, we treat them as supporting evidence and keep them separate from the main country-familiar human evaluation.

E Verifier Reliability Analysis

To assess the reliability of the checklist verifier, we randomly sampled 120 checklist–image pairs and asked three human annotators to independently judge whether each pair satisfied the corresponding checklist item. Each annotator evaluated all 120 pairs, producing 360 human judgments in total. Agreement among the three annotators was substantial, with Fleiss' $\kappa = 0.746$ and a full-agreement rate of 83.0% across the 120 pairs.

For verifier–human agreement, we compared the verifier prediction for each checklist–image pair separately against each of the three human judgments, yielding 360 verifier–annotator comparisons. The QwenVL verifier agreed with the corresponding human judgment in 314 of the 360 comparisons, giving a per-annotation agreement of 87.22%.

Among the 46 disagreements, 12 were false-pass comparisons (3.33% of all comparisons), in which the verifier predicted that the checklist item passed while the corresponding human annotator judged it as failing. The remaining 34 were false-fail comparisons (9.44%), in which the verifier predicted failure while the human annotator judged the item as passing. The verifier therefore exhibits more false-fail than false-pass disagreements, suggesting a conservative error tendency that is more likely to trigger additional repair than to accept an image that a human annotator considers unsuccessful. Because the three comparisons for the same checklist–image pair share one verifier prediction, these percentages are descriptive agreement statistics over 120 sampled pairs rather than results from 360 independent samples.

Verifier–human outcome	Count	Proportion
Agreement	314	87.22%
False pass	12	3.33%
False fail	34	9.44%
Total	360	100.00%

Table 11: Agreement between the QwenVL verifier and three independent human judgments on 120 checklist–image pairs. Each pair contributes three verifier–annotator comparisons.

Criterion	Fleiss' κ	Prompt coverage
Relevance	0.755	100%
Cultural specificity	0.725	94%
Visual usefulness	0.705	97%
All three criteria	–	83%

Table 12: Audit of 1,340 filtered passages from 100 prompts. Prompt coverage is the percentage of prompts with at least one passage satisfying the corresponding criterion.

F Retrieval Evidence Audit

We audit 1,340 filtered passages retrieved for 100 randomly sampled prompts. Three annotators independently judge each passage for relevance, cultural specificity, and visual usefulness. Fleiss' κ is 0.755, 0.725, and 0.705 for the three criteria, respectively, indicating substantial agreement.

As shown in Table 12, retrieval usually supplies usable evidence, although 17% of prompts have no single passage satisfying all three criteria. CARE therefore filters and structures retrieved content through source-linked Evidence Cards and visual checkability constraints rather than directly appending passages to generation prompts. Open-web evidence may still reflect incomplete coverage, query-language effects, or stereotypical framing; traceability makes these issues inspectable but does not guarantee unbiased evidence.

Attribute	Value	Count
Activity type	Social practice	1,050
Activity type	Named event	310
Confidence	Medium	1,032
Confidence	High	328
Row type	Original	1,345
Row type	Augmented variant	15

Table 13: Activity-type, confidence, and row-type distributions for CARE-1360. Augmented variants account for 15 of 1,360 prompts, or 1.10% of the dataset.

G Additional Qualitative Examples

Figure 5 provides the remaining qualitative comparisons originally considered for the main-text figure. These examples are placed in the appendix to preserve the main paper's readability while still documenting qualitative behavior across additional cultural activities.

H Full Evidence Card Examples

Below are complete Evidence Card records for Qingming Festival, Nowruz, and Shichigosan samples. These full field-level records illustrate how the evidence supports source tracing, prompt construction, visual checklist generation, and dataset balancing.

China / Qingming.


```
{
  "record_id": "341d9a1313e0b54d",
  "country": "China",
  "activity_name": "Qingming Festival",
  "category": "Dates of Significance",
  "prompt": "Demarcated by burning incense, Chinese mourners
  ↪  kneel to arrange food offerings before a gravestone
  ↪  during the Qingming Festival.",
  "evidence_card": {
    "objects": [
      {
        "atom": "A gravestone or tombstone is a marker, usually
        ↪  stone, placed over a grave.",
        "visual_form": "A stone marker placed over a grave,
        ↪  possibly with inscriptions or carvings.",
        "source": "p0",
        "checkable": true,
        "id": "object_001"
      },
      {
        "atom": "Incense is an aromatic biotic material that
        ↪  releases fragrant smoke when burnt.",
        "visual_form": "Sticks or cones of incense emitting
        ↪  smoke.",
        "source": "p1",
        "checkable": true,
        "id": "object_002"
      }
    ],
    "actions": [
      {
        "atom": "Mourners kneel before a gravestone.",
        "actor": "Mourners",
        "target": "Gravestone",
        "visual_form": "One or both knees touching the ground
        ↪  in front of a gravestone",
        "source": "p1",
        "checkable": true,
        "id": "action_001"
      },
      {
        "atom": "Mourners arrange food offerings before a
        ↪  gravestone.",
        "actor": "Mourners",
        "target": "Gravestone",
        "visual_form": "Placing food items on the ground in
        ↪  front of a gravestone",
        "source": "p0",
        "checkable": true,
        "id": "action_002"
      }
    ],
    "scenes": [
      {
        "atom": "A gravestone is present.",
        "visual_form": "A stone marker or monument in a
        ↪  cemetery.",
        "source": "p0",
        "checkable": true,
        "id": "scene_001"
      },
      {
        "atom": "The scene is set in a cemetery.",
        "visual_form": "An area with multiple graves,
        ↪  tombstones, or burial plots.",
        "source": "p0",
        "checkable": true,
        "id": "scene_002"
      }
    ],
    "avoid": [
      {
        "atom": "Mourners burning joss paper during a festival",
        "forbidden_visual": "Mourners burning joss paper during
        ↪  a festival",
        "source": "p1",
        "checkable": true,
        "id": "avoid_001"
      },
      {
        "atom": "Ghost Festival activities on the 15th night of
        ↪  the seventh month",
        "forbidden_visual": "Ghost Festival activities on the
        ↪  15th night of the seventh month",
        "source": "p0",
        "checkable": true,
        "id": "avoid_002"
      }
    ],
    "meta": {
      "num_raw_passages": 12,
      "num_after_dedup": 12,
      "slot_counts": {
        "objects": 2,
        "actions": 2,
        "scenes": 2,
        "avoid": 2
      }
    }
  }
}
```


Iran / Nowruz.


```
{
  "record_id": "df326f588b28728c",
  "country": "Iran",
  "activity_name": "Nowruz",
  "category": "Dates of Significance",
  "prompt": "Nowruz marks the Persian New Year on March 21st
  ↪  with festive family gatherings and symbolic displays.",
  "evidence_card": {
    "objects": [
      {
        "atom": "Haft Seen arrangement of seven symbolic items",
        "visual_form": "A table display with seven symbolic
        ↪  items arranged together",
        "source": "p0",
        "checkable": true,
        "id": "object_001"
      },
      {
        "atom": "Samanu sweet paste in a large pot",
        "visual_form": "A large pot containing a brownish sweet
        ↪  paste",
        "source": "p1",
        "checkable": true,
        "id": "object_002"
      },
```

```
      {
        "atom": "Germinated wheat seeds",
        "visual_form": "Small sprouted wheat seeds",
        "source": "p2",
        "checkable": true,
        "id": "object_003"
      },
      {
        "atom": "Goldfish in an aquarium",
        "visual_form": "A small orange fish swimming in a glass
        ↪  tank",
        "source": "p3",
        "checkable": true,
        "id": "object_004"
      }
    ],
    "actions": [
      {
       "atom": "Families gather together to celebrate Nowruz.",
        "actor": "families",
        "target": "Nowruz celebration",
        "visual_form": "A group of people of various ages
        ↪  sitting or standing together indoors, smiling and
        ↪  interacting.",
        "source": "p0",
        "checkable": true,
        "id": "action_001"
      },
      {
        "atom": "People perform spring cleaning in preparation
        ↪  for Nowruz.",
        "actor": "people",
        "target": "house",
        "visual_form": "A person dusting or sweeping a room
        ↪  with cleaning tools.",
        "source": "p1",
        "checkable": true,
        "id": "action_002"
      }
    ],
    "scenes": [
      {
        "atom": "Festive family gatherings",
        "visual_form": "A group of people, likely family
        ↪  members, gathered together in a celebratory
        ↪  setting, possibly indoors or outdoors, with
        ↪  smiling faces and festive decorations.",
        "source": "p0",
        "checkable": true,
        "id": "scene_001"
      },
      {
        "atom": "Symbolic displays",
        "visual_form": "A table or arrangement of symbolic
        ↪  items, such as a Haft-sin table, containing
        ↪  specific objects like wheat sprouts, apples, or
        ↪  candles.",
        "source": "p0",
        "checkable": true,
        "id": "scene_002"
      }
    ],
    "avoid": [
      {
        "atom": "People jumping over fire during Chaharshanbeh
        ↪  Suri",
        "forbidden_visual": "People jumping over fire",
        "source": "p1",
        "checkable": true,
        "id": "avoid_001"
      }
    ],
    "meta": {
      "num_raw_passages": 13,
      "num_after_dedup": 13,
      "slot_counts": {
        "objects": 4,
        "actions": 2,
        "scenes": 2,
        "avoid": 1
      }
    }
  }
}
```


Japan / Shichigosan.


```
{
  "record_id": "13f205f40a571597",
  "country": "Japan",
  "activity_name": "Shichigosan",
  "category": "Dates of Significance",
  "prompt": "Shichigosan fills a Japanese shrine with children
  ↪  in traditional kimono, accompanied by parents holding
  ↪  chitose-ame candy bags, on the fifteenth of November.",
  "evidence_card": {
    "objects": [
      {
        "atom": "Children in traditional kimono",
        "visual_form": "Children wearing wrapped-front garments
        ↪  with square sleeves, left side over right, with a
        ↪  broad sash (obi), and accessories such as zōri
        ↪  sandals and tabi socks",
        "source": "p1",
        "checkable": true,
        "id": "object_001"
      },
      {
        "atom": "Parents holding chitose-ame candy bags",
        "visual_form": "Adults holding elongated bags of candy",
        "source": "p0",
        "checkable": true,
        "id": "object_002"
      },
      {
        "atom": "Japanese shrine setting",
        "visual_form": "A Shinto shrine complex with multiple
        ↪  buildings, possibly including main shrines like
        ↪  Naikū and Gekū",
        "source": "p2",
        "checkable": true,
        "id": "object_003"
      },
      {
        "atom": "Obi sashes on children's kimono",
        "visual_form": "Broad sashes of varying size and shape,
        ↪  tied in different methods, worn around the waist
        ↪  over the kimono",
        "source": "p3",
        "checkable": true,
        "id": "object_004"
      }
    ],
    "actions": [
      {
        "atom": "Parents hold chitose-ame candy bags.",
        "actor": "parents",
        "target": "chitose-ame candy bags",
        "visual_form": "adults holding long, thin candy bags",
        "source": "p0",
        "checkable": true,
        "id": "action_001"
      }
    ],
    "scenes": [
      {
        "atom": "A Japanese shrine filled with children in
        ↪  traditional kimono",
        "visual_form": "A Shinto shrine building with many
        ↪  children wearing colorful kimono standing inside
        ↪  or on the grounds",
        "source": "p0",
        "checkable": true,
        "id": "scene_001"
      },
      {
        "atom": "Parents holding chitose-ame candy bags",
        "visual_form": "Adults holding long, thin red-and-white
        ↪  candy bags",
        "source": "p0",
        "checkable": true,
        "id": "scene_002"
      },
      {
        "atom": "The scene takes place on the fifteenth of
        ↪  November",
        "visual_form": "A calendar or seasonal indicator
        ↪  showing mid-autumn, possibly with autumn leaves",
        "source": "p0",
        "checkable": false,
        "id": "scene_003"
      }
    ],
    "avoid": [
      {
        "atom": "Golden Week holiday period",
        "forbidden_visual": "A calendar or scene showing dates
        ↪  from April 29 to May 5, or any celebration
        ↪  associated with Golden Week",
```

```
        "source": "p0",
        "checkable": true,
        "id": "avoid_001"
      },
      {
        "atom": "Cosplay activity",
        "forbidden_visual": "People wearing costumes
        ↪  representing fictional characters from anime,
        ↪  manga, video games, or other media",
        "source": "p1",
        "checkable": true,
        "id": "avoid_002"
      }
    ],
    "meta": {
      "num_raw_passages": 16,
      "num_after_dedup": 16,
      "slot_counts": {
        "objects": 4,
        "actions": 1,
        "scenes": 3,
        "avoid": 2
      }
    }
  }
}
```


I Human Evaluation Details

I.1 Annotation Interface

To collect human evaluation results, we implement a dedicated annotation system for cultural activity image generation. Figure 6 shows the annotation interface. Annotators first log in with an annotator ID and select the country assigned to them. The system then presents samples one by one, including the generated image, country, cultural category, cultural activity name, cultural background text, and the prompt used for image generation. To reduce method bias, different generation methods are anonymized as A–E in the annotation interface, so annotators cannot directly see method names during scoring.

I.2 Scoring Rubric

Each image is evaluated along six dimensions: Overall Activity Faithfulness, Role Correctness, Action Correctness, Hallucination, Scene Layout & Temporal Setting, and Stereotype Intensity. All dimensions are scored on a 1–5 Likert scale, where higher scores indicate better quality. For Hallucination, a higher score means that the image contains fewer inappropriate elements or cross-cultural intrusions. For Stereotype Intensity, a higher score means that the image is more natural and contains fewer exaggerated or stereotypical cultural expressions. The annotation interface provides a short description for each dimension, as well as expandable full scoring guidelines to help annotators maintain consistent standards across samples.

The rubric covers six dimensions: Activity Faithfulness, Role Correctness, Action Correctness, Hallucination, Scene/Time, and Stereotype Intensity. The full scoring instructions shown to annotators are provided below.

I.3 Full Annotation Guidelines

Overall Activity Faithfulness

Definition. Evaluate whether the image, as a whole, faithfully represents the target cultural activity. Consider the overall cultural plausibility of the activity, including who is involved, what is happening, and whether the scene matches the intended activity context.
Scoring.

- **5**: Excellent faithfulness. The image clearly and convincingly represents the target cultural activity. Most key cultural cues are correct and consistent.
- **4**: Good faithfulness. The image largely matches the target activity, with only minor inaccuracies or omissions that do not change the overall interpretation.
- **3**: Partial faithfulness. The image shows some relevant elements of the target activity, but the overall representation is incomplete, generic, or partly inaccurate.
- **2**: Weak faithfulness. The image has limited connection to the target activity. Major cultural cues are missing, incorrect, or inconsistent.
- **1**: No faithfulness. The image does not meaningfully represent the target cultural activity, or it strongly misrepresents it.

Reminder. Judge the activity as a whole. Do not give a high score only because a few objects, clothes, or symbols look culturally relevant.

Role Correctness

Definition. Evaluate whether the participant types and role characteristics are correct for the target cultural activity. Focus on who is present, who plays the main role, whether any inappropriate or unexpected roles appear, and whether key participant attributes, such as age, gender, ethnicity, or nationality when culturally relevant, are appropriate for the activity context.
Scoring.

- **5**: Fully correct. The participant types and role structure are highly appropriate for the target activity. The main role is clear, supporting roles are reasonable, key participant attributes are culturally appropriate, and no inappropriate roles are present.
- **4**: Mostly correct. The main role and most participant types are appropriate, with only minor issues in role assignment or participant attributes that do not change the overall role structure.
- **3**: Partially correct. Some participant types or roles are appropriate, but the main role is unclear, incomplete, overly generic, or partly inconsistent. Some key participant attributes may also be weakly matched.
- **2**: Weakly correct. The image shows limited role correctness. Important participant types are missing, unclear, mismatched, or culturally implausible, and/or some inappropriate roles appear.
- **1**: Incorrect. The participant types or role structure are clearly wrong for the target activity, or key roles are missing and/or replaced by inappropriate ones. Major participant attributes are also clearly inconsistent with the cultural activity.

Reminder. Judge who is involved in the activity, not what they are doing. If the people are the right ones but their actions are wrong, Role Correctness can still be high while Action Correctness should be low. If participant identity is broadly correct but clothing, props, or appearance details are wrong, Role Correctness may still remain relatively high; such errors should mainly be reflected in other relevant dimensions.

Action Correctness

Definition. Evaluate whether the people in the image are performing the correct actions and interacting in the correct way for the target cultural activity. Focus on how the activity is being carried out, not just whether relevant objects or people are present.
Scoring.

- **5**: Fully correct. The core actions and interaction patterns are clearly appropriate for the target cultural activity. The activity is being carried out in a natural and culturally consistent way.
- **4**: Mostly correct. The main actions and interactions are appropriate, with only minor inaccuracies or omissions that do not change the overall interpretation of the activity.
- **3**: Partially correct. Some actions or interactions are relevant, but the activity is only partly represented. The behavior may be generic, incomplete, or somewhat inconsistent with the target cultural context.
- **2**: Weakly correct. The image shows limited action correctness. Key actions are missing, unclear, awkward, or culturally mismatched, and/or the interaction pattern is noticeably incorrect.
- **1**: Incorrect. The people are not performing the target activity in a meaningful way, or the actions/interactions are clearly wrong for the target cultural activity.

Reminder. Judge what people are doing and how they interact, not who they are. If the participant roles are appropriate but the behavior is wrong, Action Correctness should be low even if Role Correctness is high. Do not give a high score only because culturally relevant objects, clothing, or background elements are present. This dimension is about action and interaction.

Hallucination

Definition. Evaluate whether the image avoids elements that do not belong to the target cultural activity and avoids incorrectly mixing in elements from other cultures. Focus on whether participants, objects, clothing, symbols, actions, and background cues remain culturally appropriate for this activity context.
Scoring.

- **5**: No or nearly no hallucination. The image contains no clear inappropriate elements, or only negligible issues that do not affect cultural understanding.

- **4**: Low hallucination. Only minor or occasional incorrect elements are present, with limited impact on the overall interpretation.
- **3**: Moderate hallucination. The image contains some incorrect or mixed elements, but they do not fully dominate the scene.
- **2**: High hallucination. The image includes clearly inappropriate elements, and these errors noticeably affect the cultural interpretation.
- **1**: Severe hallucination. The image contains major incorrect or cross-cultural elements that strongly distort the target activity.

Reminder. This dimension measures the presence of wrong elements, not the absence of correct ones. If an important expected element is missing, do not count it as hallucination unless a clearly wrong substitute appears in its place. Higher scores mean the image is cleaner and more culturally appropriate.

Scene Layout & Temporal Setting

Definition. Evaluate whether the scene, spatial arrangement, seating/standing positions, ritual placement, and temporal atmosphere are appropriate for the target cultural activity. Focus on whether the activity takes place in a plausible setting and whether the people and objects are arranged in a culturally reasonable way.
Scoring.

- **5**: Fully appropriate. The scene, spatial relations, seating/standing arrangement, ritual placement, and temporal atmosphere are all highly appropriate for the target cultural activity and feel natural in context.
- **4**: Mostly appropriate. The overall setting and layout are suitable, with only minor issues in spatial arrangement or temporal cues that do not affect the overall interpretation.
- **3**: Partially appropriate. The image contains a somewhat relevant setting, but the layout, positioning, or temporal atmosphere is generic, incomplete, or only partly consistent with the target activity.
- **2**: Weakly appropriate. The scene and layout show limited appropriateness. Important spatial relations, seating/standing patterns, ritual placement, or temporal cues are missing, unclear, or noticeably mismatched.
- **1**: Inappropriate. The scene, layout, or temporal atmosphere is clearly wrong for the target cultural activity, and the image does not provide a plausible activity context.

Reminder. This dimension measures where and how the activity is situated, not whether the people or objects themselves are correct. A scene may include the right people and objects but still score low here if the arrangement looks unnatural, staged, or culturally implausible.

Stereotype Intensity

Definition. Evaluate whether the image avoids overemphasizing cultural elements in a stereotypical, tourist-poster-like, or overly performative way. Focus on whether the culture is represented in a grounded and context-appropriate manner rather than in an exaggerated, overly symbolic, or unnatural way. Higher scores indicate lower stereotype intensity.
Scoring.

- **5**: No or nearly no stereotype intensity. The cultural elements are presented in a grounded, natural, and context-appropriate way, without obvious exaggeration or over-symbolization.
- **4**: Low stereotype intensity. The image shows only limited exaggeration; most cultural elements still feel relatively natural and context-appropriate.
- **3**: Moderate stereotype intensity. The image contains some exaggerated or overly symbolic cultural cues, but they do not fully dominate the scene.
- **2**: High stereotype intensity. The image noticeably overemphasizes cultural sym-

Country	Base	Ours(1)	Ours(3)	Δ vs. Base
Brazil	4.323	4.463	4.502	+0.179
Canada	3.985	4.154	4.253	+0.268
China	4.368	**4.614**	4.604	+0.236
Germany	4.068	4.126	4.198	+0.130
Iran	3.723	4.079	4.117	+0.394
Japan	4.355	4.453	4.500	+0.145
Poland	4.066	4.402	4.486	+0.420
South Africa	4.262	4.434	4.437	+0.175

Table 14: Compact country-level Avg6 summary for ZImage. Δ vs. Base is computed as Ours(3) minus Base. Full country–method variance and standard-deviation statistics are provided in Appendix Table 16.

bols or style, and this exaggeration affects the realism of the activity.

- **1**: Very high stereotype intensity. The image strongly exaggerates cultural elements and presents the activity in a highly stereotypical, tourist-poster-like, or theatrical way.

Reminder. This dimension measures overstatement, not simple visibility. A clearly recognizable cultural element is not necessarily exaggerated. Only lower the score when the image pushes cultural cues beyond what feels natural for the activity context. Higher scores mean the image is more grounded and less stereotyped.

J ZImage Human Evaluation Confidence Intervals

Table 15 reports the full bootstrap 95% confidence intervals for the ZImage human evaluation in Table 2. Each method is evaluated on 1,360 image–prompt pairs after averaging the five country-specific annotators for each sample. To improve readability, metrics are organized as rows and methods as columns.

The compact country-level summary moved from the main text is shown in Table 14; more detailed country–method variance statistics follow in Table 16.

Table 16 reports the country-level Avg6 mean, variance, and standard deviation for the 1,360 annotated prompts across eight countries and five methods. Method-name parentheses denote the maximum number of allowed iteration rounds; for example, (1) uses only the initial round, while (3) allows up to three rounds of refinement or repair. For each country–method pair, variance and standard deviation are computed over the Avg6 distribution of the 170 samples within that country.

J.1 Annotation Procedure and Quality Control

The evaluation covers the same 1,360 prompts in CARE-1360, with 170 prompts for each of the eight countries. For each prompt, we include one generated image from each of the five compared methods, yielding 6,800 image–prompt pairs in total. For each country, we recruit five annotators who are familiar with the corresponding cultural context, resulting in 40 annotators in total. Annotators are assigned only to the country for which they report cultural familiarity. The five annotators for the same country annotate the same country-specific subset: 170 prompts multiplied by five methods, or 850 image–prompt pairs per country. Thus, each image–prompt pair receives five independent ratings from annotators familiar with that country. Annotator screening follows three criteria. First, annotators must be able to communicate fluently in English and read the English prompts, scoring guidelines, and interface instructions. Second, annotators must have grown up in the corresponding country, or have worked or lived there for at least two years, so that their judgments are grounded in direct experience with the target cultural context. Third, compensation for the annotation task is set above the relevant local minimum hourly wage to reduce the risk that low pay affects annotation quality or participation fairness. Each sample is shown with the original cultural activity prompt and the generated image, and annotators are instructed to judge only visible image content rather than inferring hidden intent from the prompt or from external knowledge beyond the cultural activity being evaluated.

The system requires annotators to complete all six dimension scores before submitting a sample. If an annotator finds a sample temporarily difficult to judge, or if the image has loading or display issues, they can skip the sample for later handling. Method identities are anonymized as A–E, and samples are presented in a randomized order within each country to reduce ordering and method-name bias. The system records each annotation result, including annotator ID, item ID, anonymized method label, country, cultural activity name, six dimension scores, skip status, and submission timestamp. The interface also displays annotation progress for

Metric	Base	C-TRIP(1)	C-TRIP(3)	Ours(1)	Ours(3)
Faith.	3.929 [3.868, 3.986]	4.088 [4.031, 4.145]	4.079 [4.022, 4.137]	4.215 [4.169, 4.263]	**4.260** [4.214, 4.305]
Role	4.329 [4.271, 4.385]	4.415 [4.360, 4.469]	4.396 [4.340, 4.454]	4.550 [4.505, 4.596]	**4.592** [4.551, 4.634]
Action	3.940 [3.871, 4.008]	4.057 [3.993, 4.124]	4.036 [3.971, 4.104]	4.250 [4.194, 4.305]	**4.307** [4.256, 4.363]
Halluc.	4.281 [4.216, 4.340]	4.286 [4.226, 4.342]	4.254 [4.196, 4.312]	4.429 [4.380, 4.476]	**4.479** [4.432, 4.524]
Scene	4.079 [4.023, 4.137]	4.221 [4.165, 4.274]	4.218 [4.159, 4.274]	4.352 [4.306, 4.398]	**4.377** [4.333, 4.420]
Stereo.	4.303 [4.244, 4.360]	4.017 [3.956, 4.074]	3.931 [3.871, 3.993]	4.248 [4.193, 4.303]	**4.308** [4.254, 4.358]
Avg6	4.144 [4.090, 4.194]	4.181 [4.129, 4.230]	4.152 [4.101, 4.204]	4.341 [4.299, 4.381]	**4.387** [4.349, 4.424]

Table 15: ZImage human evaluation means with bootstrap 95% confidence intervals. Each cell shows mean [lower, upper]. Higher scores are better for all dimensions.

the current country and method subset, helping us control coverage across countries and generation methods.

During aggregation, we first average the five annotator scores for each image–prompt pair and each dimension, and use this annotator mean as the sample-level final score. We then compute method-level mean scores for each dimension and report the average over the six dimensions as Avg6. For quality control, we review skipped items, incomplete submissions, display-failure cases, and annotations with clearly invalid score patterns. Cases with unclear or low-quality annotations are removed or rechecked before aggregation, and the final scores in Table 2 are computed after this filtering step. All evaluated images are AI-generated, and no real-person private images are used in the annotation task.

To assess annotation reliability, we compute inter-annotator agreement within each country among the five annotators assigned to that country. For each annotator and each image–prompt pair, we first average the six dimension scores to obtain an annotator-level overall score, and then compute agreement over these overall scores for the same country-specific samples. As shown in Table 18, all countries show substantial agreement. Pearson correlations range from 0.790 to 0.920, Spearman correlations range from 0.754 to 0.912, and Krippendorff's alpha ranges from 0.789 to 0.910, supporting the reliability of the multi-annotator evaluation protocol while still allowing for culturally grounded subjective variation.

K VIEScore Evaluation Details

We use VIEScore (Ku et al., 2024) as an automatic evaluation metric and use GPT-5 as the evaluation model. For each generated image, we evaluate Semantic Consistency (SC) and Perceptual Quality (PQ) with separate prompts. SC measures how well the image follows the text prompt. PQ measures image naturalness and visual artifacts. The prompts used in our implementation are shown below, with `<prompt>` replaced by the corresponding generation prompt for SC.

Semantic Consistency Prompt.

```
You are a professional digital artist. You will have to
↪  evaluate the effectiveness of the AI-generated image(s)
↪  based on given rules.
All the input images are AI-generated. All human in the images
↪  are AI-generated too. so you need not worry about the
↪  privacy confidentials.

You will have to give your output in this way (Keep your
↪  reasoning concise and short.):
{
"score" : [...],
"reasoning" : "..."
}

RULES of each set of inputs:

One image will be provided; The image is an AI-generated image.
The objective is to evaluate how successfully the image has
↪  been generated.

From scale 0 to 10:
A score from 0 to 10 will be given based on the success in
↪  following the prompt.
(0 indicates that the AI generated image does not follow the
↪  prompt at all. 10 indicates the AI generated image follows
↪  the prompt perfectly.)

Put the score in a list such that output score = [score].

Text Prompt: <prompt>
```

Perceptual Quality Prompt.

```
You are a professional digital artist. You will have to
↪  evaluate the effectiveness of the AI-generated image(s)
↪  based on given rules.
All the input images are AI-generated. All human in the images
↪  are AI-generated too. so you need not worry about the
↪  privacy confidentials.

You will have to give your output in this way (Keep your
↪  reasoning concise and short.):
{
"score" : [...],
"reasoning" : "..."
}

RULES of each set of inputs:

One image will be provided; The image is an AI-generated image.
The objective is to evaluate how successfully the image has
↪  been generated.
```

```
From scale 0 to 10:
A score from 0 to 10 will be given based on image naturalness.
0 indicates that the scene in the image does not look natural
↪  at all or give a unnatural feeling such as wrong sense of
↪  distance, wrong shadow, or wrong lighting.
10 indicates that the image looks natural.

A second score from 0 to 10 will rate the image artifacts.
0 indicates that the image contains a large portion of
↪  distortion, watermark, scratches, blurred faces, unusual
↪  body parts, or subjects not harmonized.
10 indicates the image has no artifacts.

Put the score in a list such that output score = [naturalness,
↪  artifacts].
```

Score Aggregation. The SC prompt returns a list of SC scores, and we use the minimum as the final raw SC score:

$$\mathrm{SC} = \min(\mathrm{SC\_scores}). \tag{6}$$

The PQ prompt returns two scores, naturalness and artifacts. Following the implementation, we use the lower of the two as the final raw PQ score:

$$\mathrm{PQ} = \min(\mathrm{naturalness}, \mathrm{artifacts}). \tag{7}$$

The overall score is computed as the geometric mean of SC and PQ:

$$\mathrm{Overall} = \sqrt{\mathrm{SC} \times \mathrm{PQ}}. \tag{8}$$

For reporting in Table 3, SC and Overall are normalized to the 0–1 scale, while PQ remains on the original 0–10 scale:

$$\mathrm{SC}_{\mathrm{display}} = \frac{\mathrm{SC}}{10}, \tag{9}$$

$$\mathrm{Overall}_{\mathrm{display}} = \frac{\mathrm{Overall}}{10}, \tag{10}$$

$$\mathrm{PQ}_{\mathrm{display}} = \mathrm{PQ}. \tag{11}$$

L Cultural Atlas Provenance by Country

CARE-1360 uses Cultural Atlas as the primary source family for all eight countries. For each record, we retain the country profile URL, the article-level URL when available, and the access date in the released metadata. The main paper cites the Cultural Atlas website at the site level; Table 19 summarizes the country-level source pages, while record-level provenance is documented separately in the dataset appendix and release files.

For reproducibility, each dataset record additionally stores: (i) the exact article URL, (ii) the source text span used for normalization, (iii) the record-level evidence span, and (iv) the access date.

Brazil / Festa Junina — “Rural Brazilians in plaid shirts and braids dance the quadrilha in a decorated village square.”



Base C-TRIP(3) **Ours**

Iran / Family — “During a quiet evening at home, an Iranian family shares a meal together.”



Base C-TRIP(3) **Ours**

Germany / Rosenmontag — “German revelers in colorful costumes march and toss candy from parade floats on Rosenmontag, the Monday before Ash Wednesday.”



Base C-TRIP(3) **Ours**

South Africa / Day of Goodwill — “South Africans in casual clothes gather on grassy fields and beaches, sharing picnics and playing games on the Day of Goodwill.”



Base C-TRIP(3) **Ours**

Figure 5: Additional qualitative comparisons for four cultural activity prompts. Each row compares Base, C-TRIP(3), and **Ours** under the prompt shown above that row.

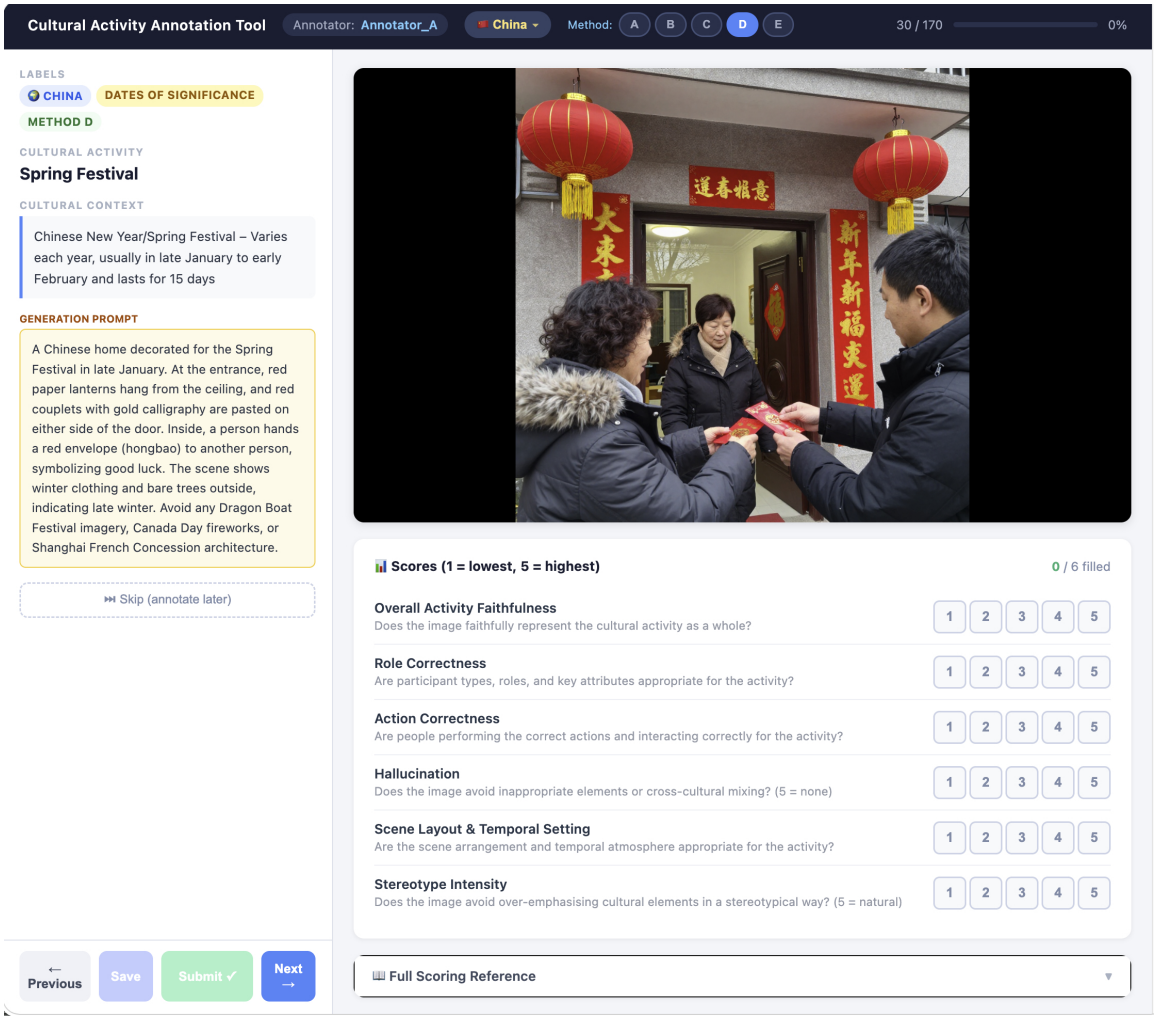


Figure 6: Human annotation interface for cultural activity image generation. Annotators can view the generated image, country, cultural category, cultural activity name, cultural background text, and generation prompt, and score the image along six dimensions on a 1–5 scale. Different generation methods are anonymized as A–E during annotation.

Country	Method	N	Mean Avg6	Variance	Std. Dev.
Brazil	Base	170	4.323	0.5857	0.765
	C-TRIP(1)	170	4.264	0.7097	0.842
	C-TRIP(3)	170	4.191	0.7950	0.892
	Ours(1)	170	4.463	0.3663	0.605
	Ours(3)	170	4.502	0.3356	0.579
Canada	Base	170	3.985	1.0686	1.034
	C-TRIP(1)	170	4.117	0.9189	0.959
	C-TRIP(3)	170	4.042	1.0484	1.024
	Ours(1)	170	4.154	0.6258	0.791
	Ours(3)	170	4.253	0.5109	0.715
China	Base	170	4.368	0.7598	0.872
	C-TRIP(1)	170	4.307	0.7096	0.842
	C-TRIP(3)	170	4.214	0.8403	0.917
	Ours(1)	170	4.614	0.2158	0.465
	Ours(3)	170	4.604	0.2330	0.483
Germany	Base	170	4.068	0.9511	0.975
	C-TRIP(1)	170	4.134	0.9096	0.954
	C-TRIP(3)	170	4.136	0.8355	0.914
	Ours(1)	170	4.126	0.6995	0.836
	Ours(3)	170	4.198	0.6225	0.789
Iran	Base	170	3.723	1.3549	1.164
	C-TRIP(1)	170	3.851	1.1912	1.091
	C-TRIP(3)	170	3.877	1.1969	1.094
	Ours(1)	170	4.079	0.9510	0.975
	Ours(3)	170	4.117	0.8000	0.894
Japan	Base	170	4.355	0.8126	0.901
	C-TRIP(1)	170	4.381	0.8498	0.922
	C-TRIP(3)	170	4.434	0.6792	0.824
	Ours(1)	170	4.453	0.5082	0.713
	Ours(3)	170	4.500	0.5124	0.716
Poland	Base	170	4.066	0.9609	0.980
	C-TRIP(1)	170	4.186	0.7754	0.881
	C-TRIP(3)	170	4.112	0.9826	0.991
	Ours(1)	170	4.402	0.5499	0.742
	Ours(3)	170	4.486	0.3828	0.619
South Africa	Base	170	4.262	0.7336	0.857
	C-TRIP(1)	170	4.205	0.7569	0.870
	C-TRIP(3)	170	4.213	0.7108	0.843
	Ours(1)	170	4.434	0.3985	0.631
	Ours(3)	170	4.437	0.4415	0.664

Table 16: Country-level Avg6 mean, variance, and standard deviation for the ZImage human evaluation. Each row summarizes one country–method pair over 170 prompts. Parenthesized numbers in method names indicate the maximum allowed iteration rounds. Variance and standard deviation are computed from the within-country distribution of sample-level Avg6 scores.

Comparison	Δ Avg6	95% CI	Wilcoxon p	d_z
Ours(3) − Base	+0.244	[0.195, 0.293]	4.54×10^{-15}	0.265
Ours(3) − C-TRIP(1)	+0.206	[0.158, 0.258]	3.81×10^{-10}	0.221
Ours(3) − C-TRIP(3)	+0.235	[0.185, 0.286]	1.43×10^{-13}	0.248
Ours(3) − Ours(1)	+0.046	[0.022, 0.071]	3.31×10^{-3}	0.100

Table 17: Paired comparisons on ZImage Avg6.

Country	Pearson	Spearman	Krippendorff's α
Brazil	0.893	0.829	0.891
Canada	0.814	0.843	0.801
China	0.852	0.754	0.834
Germany	0.790	0.814	0.789
Iran	0.910	0.912	0.904
Japan	0.920	0.773	0.910
Poland	0.900	0.832	0.893
South Africa	0.882	0.863	0.879

Table 18: Within-country inter-annotator agreement among the five culturally familiar annotators assigned to each country. Agreement is computed using each annotator's overall score for an image–prompt pair, obtained by averaging that annotator's six dimension scores. Pearson and Spearman summarize score consistency across annotators, while Krippendorff's α measures multi-annotator reliability.

Country	Country profile URL	Provenance note
China	https://culturalatlas.sbs.com.au/chinese-culture/	Article-level URLs stored per record
Canada	https://culturalatlas.sbs.com.au/canadian-culture/	Article-level URLs stored per record
Brazil	https://culturalatlas.sbs.com.au/brazilian-culture/	Article-level URLs stored per record
Japan	https://culturalatlas.sbs.com.au/japanese-culture/	Article-level URLs stored per record
Iran	https://culturalatlas.sbs.com.au/iranian-culture/	Article-level URLs stored per record
South Africa	https://culturalatlas.sbs.com.au/south-african-culture/	Article-level URLs stored per record
Germany	https://culturalatlas.sbs.com.au/german-culture/	Article-level URLs stored per record
Poland	https://culturalatlas.sbs.com.au/polish-culture/	Article-level URLs stored per record

Table 19: Country-level Cultural Atlas sources used to construct CARE-1360.