On the Cultural Gap in Text-to-Image Generation

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang,

Jinsong Su, Shuming Shi, Zhaopeng Tu

2023/07/05

Cross-Cultural Challenging (C3) Benchmark

Constructing the C3 Benchmark with GPT-4

- Ask GPT-4 to identify potential **error types** in T2I when generating cross-cultural images
- Ask GPT-4 to provide **representative Caption examples** based on error type
- Use the above as a seed to define templates, asking GPT-4 to **iterate and generate** more diverse and varied examples

- *Language Bias*: T2I systems that do not account for variations in regional dialects or Chinese script may generate text that is linguistically inaccurate or insensitive to Chinese language subtleties.
- *Cultural Inappropriateness*: Without an accurate understanding of Chinese cultural norms and values, a T2I generation system may generate images that are seen as inappropriate or offensive.
- *Missed Cultural Nuances*: T2I systems that lack an appreciation for the nuances of Chinese culture may generate images that are not authentic or credible.
- *Stereotyping and Counterfeit Representations*: T2I systems that rely on popular stereotypes or inaccurate depictions of Chinese culture may generate images that perpetuate damaging myths, or counterfeit representations give mistaken impressions.
- *Insufficient Diversity*: A T2I system that does not consider the diversity of China's 56 ethnic groups or pay attention to minority cultures' rich heritage may overgeneralize or oversimplify Chinese culture.

Table 1: Five seed captions for constructing benchmark.

A family enjoying a feast of traditional American fast food while sitting on a Chinese-style bamboo mat
A group of people performing a dragon dance at the opening of a new European-style cafe
A portrait of a woman wearing a beautiful qipao dress, holding a plate of hamburgers and fries
A bustling scene at a village fair, showcasing both Chinese lanterns and Western-style carnival games
An ancient Chinese temple adorned with modern neon signs advertising various global brands

T2I systems trained only on English data can make mistakes when generating images reflecting Chinese culture/element:
Language bias: T2I systems that do not account for variations in regional dialects
. . .
may overgeneralize or oversimplify Chinese culture.

Can you give five representative image captions in English that could lead a T2I generation trained only on English data make different types of mistakes above when generating images reflecting Chinese culture/element based on the examples but different from the examples below:

Please follow the format and only give me captions (the captions do not have to contain the word 'Chinese'), no other texts:
Example 1: Caption1
. . .
Example 5: Caption5

Cross-Cultural Challenging (C3) Benchmark

Evaluating Difficulty of the C3 Benchmark

- **C3+** 9,889 instances, with a sample of 500 (C3) used for manual assessment
- **Assessment Difficulty**: ask Stable Diffusion to generate images based on different datasets.
- The score was "≥3" for 78% on COCO and **57%** on **C3**. For C3, 26.2% scored 1 point.

	C^3	C^3+	COCO
Caption	500	9,889	500
Length	29.34	26.49	10.22
Object	10.76	9.81	3.65

(a) **Data Statistics**

(b) C^3 (c) C^3+ (d) **MS-COCO**

Figure 2: Statistics (a) and Word Cloud (b,c) of the C^3 benchmark and its expanded edition C^3+. "Length" and "Object" denote the average number of words and objects in each caption, respectively. We list the details of the MS-COCO Captions ("MS-COCO") benchmark for reference.

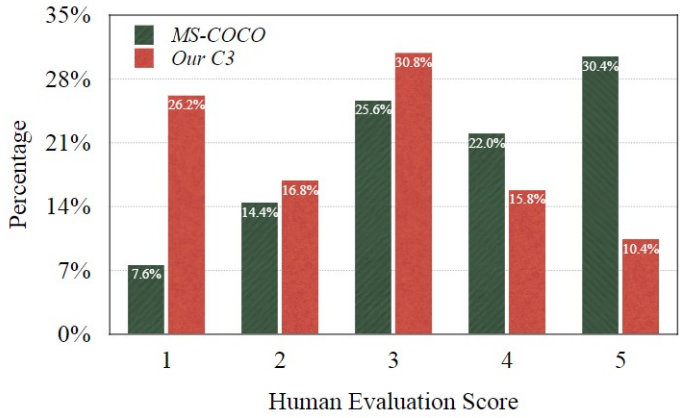
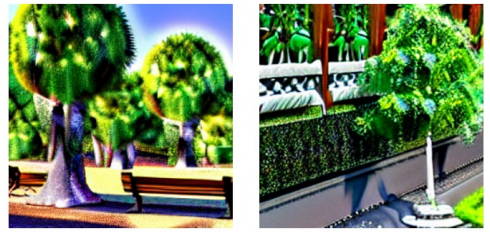


Figure 3: Human scoring results of Stable Diffusion on the widely-used MS-COCO and the proposed C^3 benchmarks. The overall human score that considers both *text-image alignment* and *fidelity* on MS-COCO and C^3 are respectively 3.53 and 2.67, suggesting that C^3 is more challenging.



(1) A park bench in the midst of a beautiful desert garden.
(2) An outdoor garden area with verdant plants and a tree.

(a) **MS-COCO Benchmark**

(1) A serene scene of a tea ceremony in a serene Chinese garden setting.
(2) A beautiful Chinese garden with a gracefully arched bridge and blooming lotus flowers.

(b) **C^3 Benchmark**

Figure 4: Example images generated by the Stable Diffusion v1-4 model on the MS-COCO and C^3 benchmarks. We highlight in red the objects missed in the generated image.

Cross-Cultural Challenging (C3) Benchmark

Human Evaluation Criteria for the C3 Benchmark

- Existing human evaluation criteria do not adequately reflect cultural differences.
- Propose **fine-grained evaluation criteria** covering image-text alignment and image fidelity, as well as characteristics (e.g. cultural appropriateness) and challenges (e.g. cross-cultural object presence and localization)

1. **Cultural Appropriateness** that examines the extent to which the generated images reflect the cultural style and context mentioned in the caption. This criterion helps to demonstrate the model's ability to capture and generate culturally relevant visual content.
2. **Object Presence** that evaluates whether the generated images contain the essential objects mentioned in the caption. This criterion ensures that the model accurately generates the cross-cultural objects in the caption.
3. **Object Localization** that assesses the correct placement and spatial arrangement of objects within the generated images, which can be challenging for the cross-cultural objects. This criterion ensures that the model maintains the context and relationships between objects as described in the caption.
4. **Semantic Consistency** that assesses the consistency between the generated images and the translated captions, ensuring that the visual content aligns with the meaning of the text. This criterion evaluates the model's ability to generate images that accurately represent the caption.
5. **Visual Aesthetics** that evaluates the overall visual appeal and composition of the generated images. This criterion considers factors such as color harmony, contrast, and image sharpness, which contribute to the perceived quality of the generated images.
6. **Cohesion** that examines the coherence and unity of the generated images. This criterion evaluates whether all elements appear natural and well-integrated, contributing to a cohesive visual scene.

Figure 1: Comparison of the original stable diffusion (left) and the stable diffusion fine-tuned on the dataset filtered by our approach (right) for generating cross-cultural images with Chinese elements based on the prompt *A garden with typical Chinese architecture and design elements*. The example clearly demonstrates that the fine-tuned system can produce higher quality images.

Table 2: Evaluation scores for the example image generated by the vanilla stable diffusion model in Figure 1 (left panel).

Criteria	S	Reasons
Cultural Appropriate	3	The specific cultural elements and styles of China can be distinguished in the image, but there are some meaningless parts.
Object Presence	3	Some objects can be seen in the image, but it is difficult to distinguish specific elements.
Object Localization	2	The temple elements in the image are not lined up correctly.
Semantic Consistency	2	The consistency between the image and the caption is poor.
Visual Aesthetics	1	Overall image quality is very poor.
Cohesion	2	Multiple elements in the image are not coherently matched.

Improving Cross-Cultural Generation

Motivation

- Typically, in-domain captions are translated into English and the **translated image-caption** pairs are used for fine-tuning diffusion models.
- A significant challenge lies in **filtering out low-quality** translated captions.
- We introduce **a new filtering method** considering multi-modal alignment: text-text, image-text alignment, and explicit object-text alignment.

Revisiting Existing Methods

- Text-Text Alignment uses reference-free metrics such as BERTscore and LaBSE.
- Image-Text Alignment uses multi-modal pre-trained vision-language models such as CLIP.
- Previous research primarily used these methods to filter in-domain data, neglecting other useful alignment information.

Improving Cross-Cultural Generation

Our Approach – Multi-Modal Alignment

- **Text-Text Alignment (S-T)** original and the translated captions
- **Image-Text Alignment (I-T)** image and the translated caption
- **Object-Text Alignment (O−T)** detected objects in the image and the translated caption

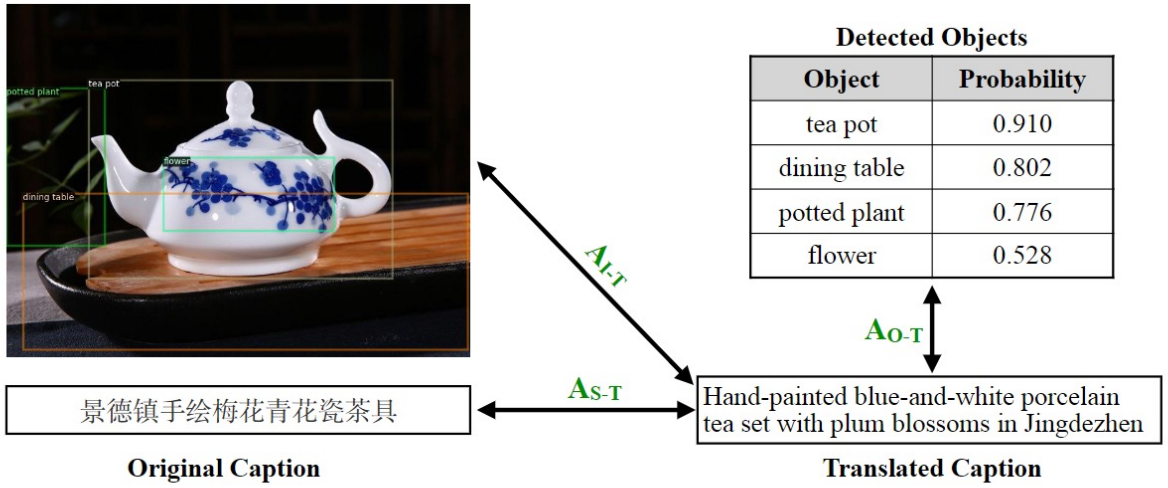


Figure 5: Framework of our filtering metric that measures the quality of the translated caption with three alignment scores: 1) A_{S-T} for aligning the original caption; 2) A_{I-T} for aligning the image; and 3) A_{O-T} for aligning the detected objects.

We follow (Zhang et al. 2019) to calculate the text-text alignment between two captions as a sum of cosine similarities between their tokens' embeddings:

$$A_{S-T} = \frac{1}{M} \sum_{\mathbf{x} \in \mathbf{H}_S} \max_{\mathbf{y} \in \mathbf{H}_T} \frac{\mathbf{x}^\top \mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} \tag{5}$$

Similarly, we calculate the other two alignment scores by:

$$A_{O-T} = \frac{1}{K} \sum_{\mathbf{o} \in \mathbf{H}_O} \max_{\mathbf{y} \in \mathbf{H}_T} \frac{\mathbf{o}^\top \mathbf{y}}{||\mathbf{o}||\,||\mathbf{y}||} \tag{6}$$

$$A_{I-T} = \max_{\mathbf{y} \in \mathbf{H}_T} \frac{\mathbf{h}_I^\top \mathbf{y}}{||\mathbf{h}_I||\,||\mathbf{y}||} \tag{7}$$

The ultimate score is a combination of the above alignments:

$$A = A_{S-T} + A_{I-T} + A_{O-T} \tag{8}$$

Experiments

Dataset

- Chinese subset (**laion2b-zh**) of the **laion2b-multi** dataset, comprising a total of **143 million** image-text pairs
- We translate all image captions **into English** using an online translation system TranSmart
- **Filter** the full laion-zh to **300K** instances with different strategies, including 1) the text-text alignment score LaBSE; 2) the image-text alignment score CLIP; 3) our multi-modal metric

Model

- We **fine-tune** the diffusion model on the filtered laion-zh dataset for **one epoch** with a batch size of 2 on 8 A100 40G GPUs.
- We use the AdamW optimizer with a learning rate of 1e-4 for all models

Experiments

Assessing the Quality of Translated Caption

- Randomly sampled 500 instances from the translated laion2b-zh data
- Ask **human annotators** to rate the quality of translated caption
  - **textual translation quality**, including adequacy, fluency and consistency;
  - **image correlation**, including image relevance, context, and cultural appropriateness.
- Score the translated captions with **different automatic metrics**
- Calculate their Pearson correlation with the **human judgements**

Table 6: Evaluation guidelines for the translated captions associated with the images.

Score	Adequacy	Fluency	Consistency
5	The translation accurately conveys the intended meaning of the original caption with no errors or inaccuracies.	The translation is very well-written, with no errors in grammar, syntax, or vocabulary that could impact understanding.	The translations are consistent in language, tone, and style, with no noticeable differences.
4	The translation accurately conveys the intended meaning of the original caption with only minor errors or inaccuracies.	The translation is well-written, with only minor errors in grammar, syntax, or vocabulary that do not impact understanding.	The translations are mostly consistent in language, tone, and style, with minor differences that are hardly noticeable.
3	The translation mostly conveys the intended meaning of the original caption, but may still have some errors or inaccuracies.	The translation is generally well-written, with only a few errors in grammar, syntax, or vocabulary that do not significantly impact understanding.	The translations are generally consistent in language, tone, and style, with only a few noticeable differences.
2	The translation partially conveys the intended meaning of the original caption, but misses some important details or nuances.	The translation is somewhat fluent, but still contains some errors in grammar, syntax, or vocabulary that may make it slightly difficult to understand.	The translations are somewhat consistent, but still contain noticeable differences in language, tone, or style that may be distracting.
1	The translation does not convey the intended meaning of the original caption at all.	The translation is poorly written, with numerous errors in grammar and syntax that make it difficult to understand.	The translations are inconsistent in language, tone, or style, making them difficult to follow.

Score	Relevance	Context	Cultural appropriateness
5	The translations are perfectly relevant to the image they describe, capturing the essence of the image and all important details in a highly engaging way.	The translations provide perfect context for the reader to understand the image and the situation in which it was taken, leaving no room for confusion.	The translations are perfectly appropriate for the target audience, demonstrating a deep understanding of the target culture.
4	The translations are highly relevant to the image they describe, capturing the essence of the image and all important details.	The translations provide highly sufficient context for the reader to understand the image and the situation in which it was taken, with only minor room for confusion or ambiguity.	The translations are highly appropriate for the target audience, with minimal cultural references or language that could be offensive or confusing.
3	The translations are somewhat relevant to the image they describe, capturing some important details but lacking in depth or engagement.	The translations provide some context for the reader to understand the image and the situation in which it was taken, but may be somewhat confusing.	The translations are somewhat appropriate for the target audience, with some cultural references or language that may be slightly offensive or confusing.
2	The translations are minimally relevant to the image they describe, lacking important details and failing to engage the reader.	The translations provide little context for the reader to understand the image and the situation in which it was taken, leaving much room for confusion or ambiguity.	The translations are minimally appropriate for the target audience, with cultural references or language that may be offensive or confusing.
1	The translations are not relevant to the image they describe, failing to capture the essence of the image and important details.	The translations provide no context for the reader to understand the image and the situation in which it was taken, causing confusion or ambiguity.	The translations are not appropriate for the target audience, with cultural references or language that is offensive or confusing.

Experiments

Assessing the Quality of Translated Caption

- Our metric **outperforms** LaBSE and CLIP in terms of correlation with human assessment scores across all standards
- A **positive correlation coefficient** indicates a strong consistency between the multi-modal alignment measure and human judgment
- Ours more effectively **captures key aspects** of the T2I generation task compared to others
- We remove object-text alignment scores to investigate **the influence within our metric**

Table 3: Pearson correlation ($p < 0.01$) with sentence-level human judgments from different perspectives. "All" denotes the overall Pearson correlation in all criteria. "$-A_{O-T}$" denotes removing the object-text alignment score A_{O-T} from our metric.

Filtering	**Textual Translation Quality**			**Image Correlation**			**All**
Metric	**Adequacy**	**Fluency**	**Consistency**	**Relevance**	**Context**	**Appropriateness**	
LaBSE	0.107	-0.033	0.194	0.167	0.215	0.125	0.129
CLIP	-0.081	-0.114	-0.092	-0.085	-0.057	-0.086	-0.086
Ours	**0.220**	**0.149**	**0.295**	**0.220**	**0.215**	0.163	**0.211**
$-A_{O-T}$	0.098	-0.050	0.185	0.158	0.211	0.115	0.119
A_{O-T}	0.210	0.161	0.274	0.200	0.186	0.148	0.197

Experiments

Performance on the C3 Benchmark

- **Importance of fine-tuning**: **all models** that were fine-tuned showed better performance than those only trained on English data
- **Necessity of Filtering Low-Quality Translations**: filtering methods with **specific metrics** outperform random sampling strategies
- **Superiority of our Metric**: by preserving **high-quality instances** for fine-tuning, the best results are obtained under all standards

Table 4: Human evaluation of the images generated by vanilla and fine-tuned diffusion models on the C^3 benchmark.

System	Presence	Localization	Appropriateness	Aesthetics	Consistency	Cohesion
Vanilla	3.66	3.50	3.61	3.06	3.39	3.17
Fine-Tuned on Chinese-Cultural Data						
Random	4.27	4.19	4.22	3.65	4.08	3.96
LaBSE	4.68	4.47	4.61	3.72	4.39	4.16
CLIP	4.66	4.54	4.56	3.87	4.38	4.12
Ours	**4.74**	**4.65**	**4.71**	**3.92**	**4.53**	**4.33**

Experiments

Performance on the C3 Benchmark

- The vanilla diffusion model struggles to produce Chinese cultural elements. This issue is significantly alleviated through model fine-tuning.
- While both CLIP and our model successfully generate all objects in the captions (e.g. *tea ceremony with an expert* and *winding pathways*, *carefully placed rocks*, and *lush vegetation*), elements in **our images** appear **more natural and integrated**

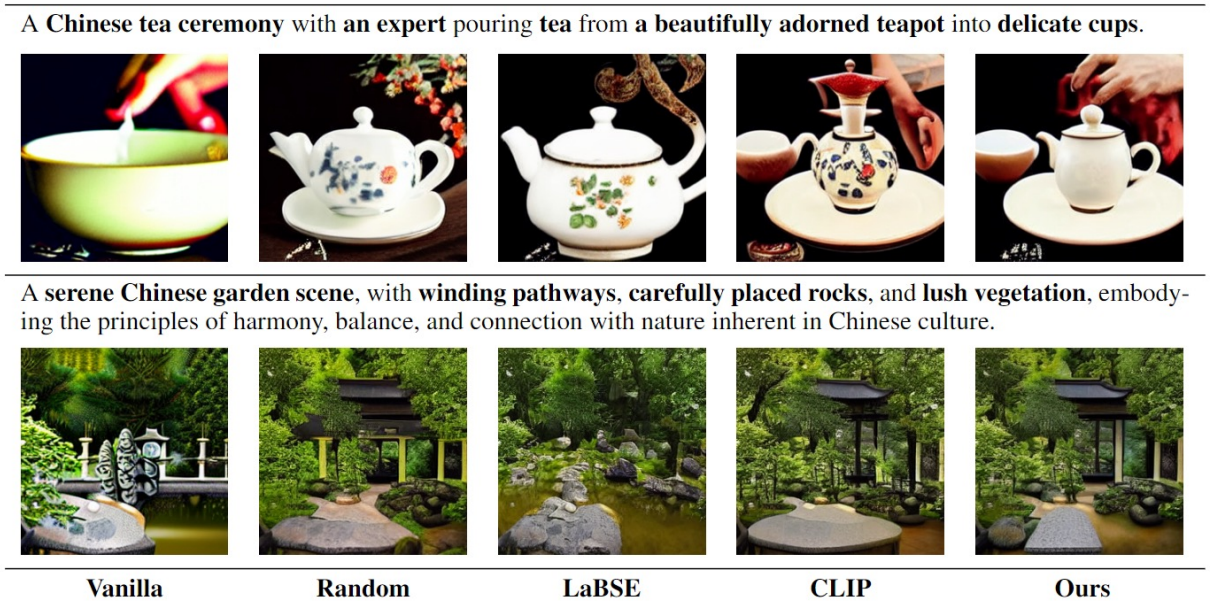


Figure 6: Example images generated by vanilla and fine-tuned diffusion models. We highlight in **bold** the objects in the caption.