ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240581

On the Cultural Gap in Text-to-Image Generation

Bingshuai Liu[a,b,1], **Longyue Wang**[a,*], **Chenyang Lyu**[a,c], **Yong Zhang**[a], **Jinsong Su**[b], **Shuming Shi**[a] and **Zhaopeng Tu**[a,*]

[a]Tencent AI Lab
[b]Xiamen University
[c]Dublin City University

Abstract. One challenge in text-to-image (T2I) generation is the inadvertent reflection of culture gaps present in the training data, which signifies the disparity in generated image quality when the cultural elements of the input text are rarely collected in the training set. Although various T2I models have shown impressive but arbitrary examples, there is no benchmark to systematically evaluate a T2I model's ability to generate cross-cultural images. To bridge the gap, we propose a Challenging Cross-Cultural (C^3) benchmark with comprehensive evaluation criteria, which can assess how well-suited a model is to a target culture. By analyzing the flawed images generated by the Stable Diffusion model on the C^3 benchmark, we find that the model often fails to generate certain cultural objects. Accordingly, we propose a novel multi-modal metric that considers object-text alignment to filter the fine-tuning data in the target culture, which is used to fine-tune a T2I model to improve cross-cultural generation. Experimental results show that our multi-modal metric provides stronger data selection performance on the C^3 benchmark than existing metrics, in which the object-text alignment is crucial. We release the benchmark, data, code, and generated images to facilitate future research on culturally diverse T2I generation.

1 Introduction

Text-to-image (T2I) generation has emerged as a significant research area in recent years, with numerous applications spanning advertising, content creation, accessibility tools, human-computer interaction, language learning, and cross-cultural communication [26]. One challenge of T2I models is the inadvertent reflection or amplification of cultural gaps present in the training data, which refer to differences in norms, values, beliefs, and practices across various cultures [21, 32]. The cultural gap in T2I generation signifies the disparity in image generation quality when the cultural elements of the input text are rarely collected in the training set. For example, in the LAION 400M dataset, the collected text-image pairs predominantly consist of English texts and images containing Western cultural elements. Consequently, given a text description featuring Eastern cultural elements, the quality of the generated image is likely to be unsatisfactory. Figure 1 shows an example. The Stable Diffusion model that is trained on the Western cultural data fails to generate satisfying Chinese cultural elements.

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.

The lack of cultural sensitivity in the generated images can manifest in the form of images that may be inappropriate, offensive, or simply irrelevant in certain cultural contexts. Therefore, addressing these cultural gaps in AI T2I models is crucial to ensure the generation of culturally appropriate and contextually relevant images for users from diverse cultural backgrounds. However, although various T2I models have shown how the cultural gap leads to flawed images with impressive but arbitrary examples, there is no benchmark to systematically evaluate a T2I model's ability to generate cross-cultural images.

To bridge the gap, we introduce a C^3 benchmark with comprehensive evaluation criteria for the target evaluation on the cross-cultural T2I generation. Given that current open-sourced T2I models are generally trained on the English data associated with Western cultural elements [26, 24], we built a evaluation set of textual prompts designed for generating images in Chinese cultural style. Specifically, we ask the powerful GPT-4 model with carefully designed context to generate the challenging prompts that can lead a T2I model to make different types of cross-cultural generation errors. We also provide a set of evaluation criteria that consider characteristics (e.g. cultural appropriateness) and challenges (e.g. cross-cultural object presence and localization) of cross-cultural T2I generation.

A promising way of improving cross-cultural generation is to fine-tune a T2I model on training data in target culture, which are generally in other non-English languages. Accordingly, the captions in the target-cultural data are translated to English with external translation

* Longyue Wang and Zhaopeng Tu are corresponding authors. Email: {vinnylywang, zptu}@tencent.com.
[1] Work done while Bingshuai Liu and Chengyang Lyu were interning at Tencent AI Lab.

systems, which may introduce translation mistakes that can affect the quality of the image-caption pairs. In response to this problem, we propose a novel multi-modal metric that considers both textual and visual elements to filter low-quality translated captions. In addition, analyses of generated images on the C^3 benchmark show that the object generation in target culture is one of the key challenges for cross-culture T2I generation. Accordingly, our multi-modal metric includes an explicit object-text alignment score to encourage that all necessary objects in the image are included in the translated caption. Empirical analysis shows that our metric correlates better with human judgement on assessing the quality of translated caption for T2I than existing metrics. Experimental results on the C^3 benchmark show that our multi-modal metric provides stronger data selection performance. In summary, **our contributions** are as follows:

- We build a benchmark with comprehensive evaluation criteria for cross-cultural T2I generation, which is more challenging than the commonly-used MS-COCO benchmark with more cross-cultural objects.
- We propose a multi-modal metric that considers both textual and visual elements to filter training data in the target culture, which produce better performance for fine-tuning a T2I model for cross-cultural generation.
- To facilitate future research on culturally diverse T2I generation, we publicly release the resources we constructed in this paper, including the C^3 benchmark, translated dataset, the filtering scripts, and generated images.

2 Related Work

In the last several years, there has been a growing interest in T2I generation. The conventional generation models are built upon generative adversarial networks (GANs) [25, 35, 36], which consists of a text encoder and an image generator. Recently, diffusion models have advanced state of the art in this field by improving image quality and diversity [24, 23, 26, 29]. Previous research on text-guided image generation mainly focused on improving the understanding of complex text descriptions [38, 28] or the quality of generated images [29]. In this work, we aim to improve the generalization of T2I models to generate images associated with cultural elements that have rarely been observed in the training data.

Another thread of research turns to enhance multilingual capabilities of T2I models, which can support non-English input captions. For example, Chen et al. [4] extent the text encoder of diffusion model with a pre-trained multilingual text encoder XLM-R. Li et al. [13] mitigated the language gap by translating English captions to other languages with neural machine translation systems.

Chen et al. [3] introduced the PaLI model, which is trained on a large multilingual mix of pre-training tasks containing 10B images and texts in over 100 languages. This model emphasizes the importance of scale in both the visual and language parts of the model and the interplay between the two.

Saxon and Wang [30] proposed a novel approach for benchmarking the multilingual parity of generative T2I systems by assessing the "conceptual coverage" of a model across different languages. They build an atomic benchmark that narrowly and reliably captures a specific characteristic – conceptual knowledge as reflected by a model's ability to reliably generate images of an object across languages. Similarly, we build a benchmark to capture another specific characteristic – cross-cultural generation as reflected by a model's ability to reliably generate cultural elements that are rarely collected in the training set.

Closely related to this work, Liu et al. [17] also concerns the cross-culture T2I problem. Our works are complementary to each other: we focus on building a comprehensive benchmark for the target evaluation on the cross-cultural T2I generation, while they aim to improving the cross-cultural performance with the prompt-augmentation and standard fine-tuning. In addition, our multi-modal alignment approach can further improve their model performance by enhancing the fine-tuning process.

3 Cross-Cultural Challenging (C^3) Benchmark

3.1 Constructing the C^3 Benchmark with GPT-4

To generate captions for creating cross-cultural and culturally diverse images, we firstly summarise several types of mistakes T2I generation systems can make if they are asked to generate such cross-cultural images, which serve as the prompt for GPT-4 to generate more challenging captions:

- *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 captions.
- *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.

Subsequently, we asked GPT-4 to provide five representative examples of image captions in English that could lead a T2I system, trained only on English data, to make different types of mistakes when generating images reflecting Chinese culture or elements, as listed in Table 1. We used the first five examples (selected and checked by humans) as seed examples to iteratively generate more diverse and different examples, which can lead to errors while generating images reflecting Chinese culture or elements. Specifically, we use the following prompt to obtain more challenging captions:

> T2I systems trained only on English data can make mistakes when generating images reflecting Chinese culture/element:
> Language bias: · · ·
> Cultural Inappropriateness: · · ·
> · · ·
>
> *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

Table 1. Five seed captions for constructing benchmark.

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

Table 2. Data Statistics of C^3 Benchmark and COCO.

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

In each iteration we randomly sample five seed examples from the generated examples as prompt examples. The collected image captions were used to construct an evaluation set for assessing the performance of T2I generation systems in generating cross-cultural and culturally diverse images. Finally, we obtain a set of 9,889 challenging captions by filtering the repetitive ones for cross-cultural T2I generation, which we name as C^3+. Since it is time-consuming and labor-intensive to manually evaluate the generated images for all the captions, we randomly sample 500 captions to form a small-scale benchmark C^3, which will serve as the testbed in the following experiments for human evaluation. The generated images for different models on the full C^3+ benchmark (without human evaluation) will also be released for future research. Table 2 and Figure 4 shows the benchmark details.

3.2 *Evaluating Difficulty of the C^3 Benchmark*

To evaluate the difficulty of the C^3 benchmark, we compare with the commonly-used COCO Captions dataset [2], which is extracted from the English data that is potentially similar in distribution with the training data of Stable Diffusion. Specifically, we sample 500 captions from the COCO data, and ask the Stable Diffusion v1.4 model to generate images based on the captions. Figure 4 shows the details of the sampled COCO Caption data. Compared with C^3, the captions in COCO contain smaller sizes of words and objects, which makes it easier for T2I generation.

For comparing the quality of the generated images on both benchmarks, we follow the common practices to ask human annotators to score the generated images from the perspectives of both the image-text alignment and image fidelity [29, 8]. Figure 2 lists the comparison results. Clearly, 78% of the generated images on COCO are rated above average ("$\geq$ 3"), while the ratio on C^3 is 57%. Specifically, 26.2% of the generated images on C^3 is rated as the lowest 1 score, which is far larger than that on COCO. Figure 3 shows some examples of generated images on the two benchmarks. The Stable Diffusion model successfully generates all objects in the MS-COCO captions. However, it fails to generate cultural objects (e.g. "a tea ceremony", "a gracefully arched bridge", and "blooming lotus flowers")

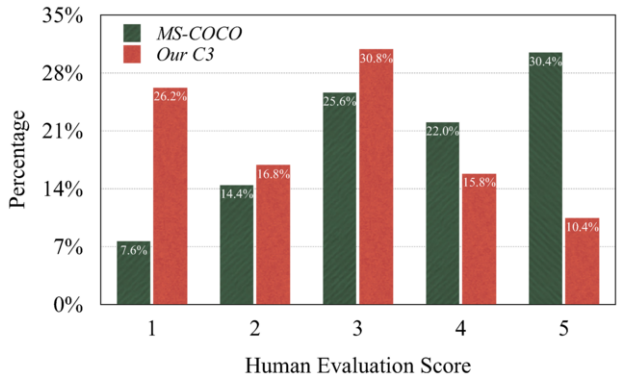
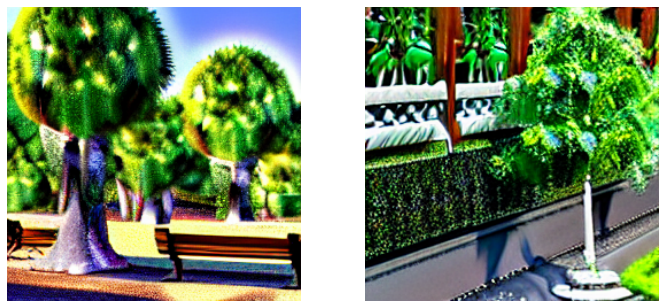
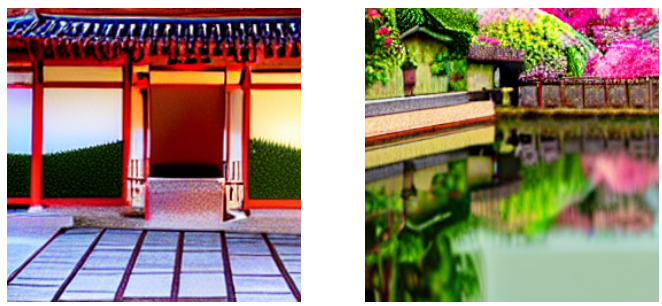


Figure 2. Human scoring results of Stable Diffusion on the widely-used MS-COCO and the proposed C^3 benchmarks.



(1) **left:** A park bench in the midst of a beautiful desert garden.
(2) **right:** An outdoor garden area with verdant plants and a tree.
(a) MS-COCO Benchmark



(1) **left:** A serene scene of a tea ceremony in a serene Chinese garden setting.
(2) **right:** A beautiful Chinese garden with a gracefully arched bridge and blooming lotus flowers.
(b) C^3 Benchmark

Figure 3. 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 image.

in the C^3 captions, which are rarely observed in the training data of the diffusion model. These results demonstrate that the proposed C^3 is more challenging.

3.3 *Human Evaluation Criteria for C^3 Benchmark*

Although the metrics of image-text alignment and image fidelity are widely-used for general T2I generation, they may not be sufficient to capture the certain types of mistake in the cross-cultural scenario (e.g. cultural inappropriateness and object presence). In response to this problem, we propose a fine-grained set of criteria for the target

Figure 4. Word Cloud 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.

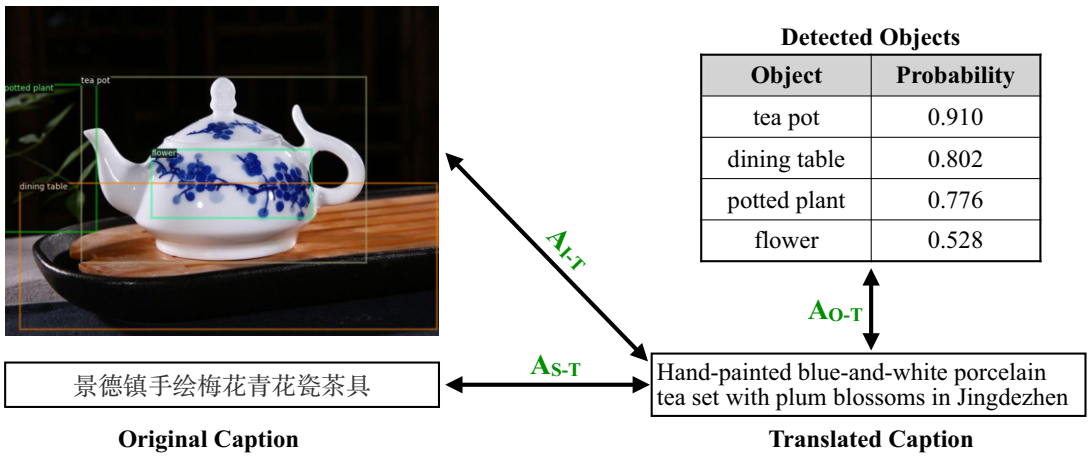


Figure 5. Framework of our filtering metric that measures the quality of the translated caption with three alignment scores: 1) A_{S-T} for aligning the original caption; 2) A_{I-T} for aligning the image; and 3) A_{O-T} for aligning the detected objects.

Table 3. 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.

evaluation on the cross-cultural T2I generation, which focuses on various aspects of cultural relevance and image quality:

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 nat-

ural and well-integrated, contributing to a cohesive visual scene.

As seen, in addition to generalizing the conventional image-text alignment (e.g. semantic consistency) and image fidelity (e.g. visual aesthetics and cohesion) criteria, we also propose several novel metrics that consider characteristics (e.g. cultural appropriateness) and challenges (e.g. cross-cultural object presence and localization) of cross-cultural T2I generation. We hope the fine-grained evaluation criteria can provide a comprehensive assessment of the generated images on the proposed C^3 benchmark. Table 3 lists an example of using the criteria to evaluate the image in Figure 1 (left panel). Table 5 in the supplementary material [16] lists the guideline of using these criteria for human evaluation.

4 Improving Cross-Cultural Generation

A promising way of improving cross-cultural T2I generation is to fine-tune the diffusion model on the in-domain data (e.g. image-text pairs of Chinese cultural in this work). Generally, the captions of the in-domain data are translated into English, and the pairs of (translated caption, image) are used to fine-tune the diffusion model. The main challenge lies in how to filter low-quality translated captions.

In this section, we first revisit existing filtering methods, which considers only either text-text alignment or image-text alignment. Inspired by recent successes on multi-modal modeling [19], we propose a novel filtering approach that considers **multi-modal alignment** including both text-text and image-text alignment, as well as explicit object-text alignment since the objects are one of the key challenges for cross-cultural T2I generation.

4.1 Revisiting Existing Methods

Text-Text Alignment Since there is no reference translation for captions of in-domain data, conventional metrics such as BLEU [20] and Meteor[1] that rely on the reference are unsuitable for evaluating the quality of the translated captions. Accordingly, researchers turn to reference-free metric such as BertScore [37], which computes a similarity score for two sentences in the same language by leveraging the pre-trained contextual embeddings from BERT. Along this direction, Feng et al. [7] propose a multilingual version – LaBSE, which can compute a similarity score for two sentences in different languages.

Image-Text Alignment Another thread of research uses multi-modal pre-trained vision-language models to measure the alignment between caption and images. One representative work is CLIP [22], which computes a similarity score for a sentence and image with a pre-trained model on a dataset of 400 million (image, text) pairs. While prior studies use only either text-text alignment or image-text alignment for filtering the in-domain data, they miss the useful information from the other alignment. In response to this problem, we propose a multi-modal alignment approach to better measure the quality of the (image, translated caption) pair.

4.2 Our Approach – Multi-Modal Alignment

As shown in Figure 5, our filtering metric consists of three types of alignment scores: 1) *Text-Text Alignment* A_{S-T} between the original and translated captions; 2) *Image-Text Alignment* A_{I-T} between the image and the translated caption; 3) *Object-Text Alignment* A_{O-T} between the detected objects in the image and the translated caption.

Formally, let $S = \{x_1, \cdots, x_M\}$ be the original non-English caption associated with the image I, $T = \{y_1, \cdots, y_N\}$ be the translated caption in English, and $O = \{o_1, \cdots, o_K\}$ be the list of the objects (listed in natural language) detected in the image I. We first encode the captions and objects with a multilingual BERT $\mathcal{E} \in \mathbb{R}^h$ [5] to the corresponding representations:

$$\mathbf{H}_S = \mathcal{E}(S), \mathbf{H}_T = \mathcal{E}(T), \mathbf{H}_O = \mathcal{E}(O) \tag{1}$$

where $\mathbf{H}_S \in \mathbb{R}^{M\times h}$, $\mathbf{H}_T \in \mathbb{R}^{N\times h}$ and $\mathbf{H}_O \in \mathbb{R}^{K\times h}$.

We encode the image I with a Vision Transformer $\mathcal{V} \in \mathbb{R}^h$ [6] into a representation vector:

$$\mathbf{h}_I = \mathcal{V}(I) \qquad \in \mathbb{R}^h \tag{2}$$

We follow [37] to calculate the text-text alignment between two captions as a sum of cosine similarities between their tokens' embeddings:

$$A_{S-T} = \frac{1}{M} \sum_{\mathbf{x}\in\mathbf{H}_S} \max_{\mathbf{y}\in\mathbf{H}_T} \frac{\mathbf{x}^\top \mathbf{y}}{||\mathbf{x}||\ ||\mathbf{y}||} \tag{3}$$

Similarly, we calculate the other two alignment scores by:

$$A_{O-T} = \frac{1}{K} \sum_{\mathbf{o}\in\mathbf{H}_O} \max_{\mathbf{y}\in\mathbf{H}_T} \frac{\mathbf{o}^\top \mathbf{y}}{||\mathbf{o}||\ ||\mathbf{y}||} \tag{4}$$

$$A_{I-T} = \max_{\mathbf{y}\in\mathbf{H}_T} \frac{\mathbf{h}_I^\top \mathbf{y}}{||\mathbf{h}_I||\ ||\mathbf{y}||} \tag{5}$$

The ultimate score is a combination of the above alignments:

$$A = A_{S-T} + A_{I-T} + A_{O-T} \tag{6}$$

The score A reflects the quality of the translated captions by considering both their textual and visual information. A higher A indicates that the translated caption has better quality with respect to the original caption, the relatedness between image and caption, and the similarity between image and caption at an object-level. Each term in A measures the translation quality from a specific aspect, thereby allowing for a faithful reflection of the overall translation quality. Practically, we followed previous work to implement the text-text alignment A_{S-T} with LaBSE and implement the image-text alignment A_{I-T} with CLIP. We use GRiT [34] to implement A_{O-T}. GRiT will detect objects in the image and output corresponding categories. We detect the objects in the images using the GRiT model with prediction probability > 0.5.

In summary, our proposed approach involves the following steps:

1. Obtaining embeddings of the original captions, translated captions and images.
2. Extracting objects in images and encode object labels.
3. Calculating text-to-text, image-to-text and object-to-text similarity scores.
4. Calculate the translation quality score as the combination of the three similarity scores.

Our approach provides a novel method for estimating the quality of translated captions in non-English T2I datasets, and has the potential to improve the performance of image generation models by incorporating data in other languages.

Table 4. 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

Table 5. 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**

4.3 Experiments

Experimental Setup We conduct experiments with the Stable Diffusion v1-4 model [26].[2] For fine-tuning the diffusion model on the Chinese cultural data, we choose the Chinese subset (*laion2b-zh*) of the *laion2b-multi* dataset[3], comprising a total of 143 million image-text pairs. We translate all image captions into English using an online translation system TranSmart [9] (https://transmart.qq.com).

We filter the full *laion-zh* to 300K instances with different strategies, including 1) the text-text alignment score **LaBSE** [7]; 2) the image-text alignment score **CLIP** [22]; 3) our multi-modal metric. 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 [18] with a learning rate of 1e-4 for all models.

Assessing the Quality of Translated Caption We randomly sampled 500 instances from the translated *laion2b-zh* data, and ask human annotators to rate the quality of translated caption from two main perspectives: 1) textual translation quality, including adequacy, fluency and consistency; and 2) image correlation, including image relevance, context, and cultural appropriateness. Table 6 in the supplementary material [16] lists the evaluation guidelines. We then scored the translated captions with different automatic metrics (e.g. LaBSE, CLIP, and Ours), and calculate their Pearson correlation with the human judgements on the above criteria. As for the details of human annotation, we recruited 5 annotators that are native Chinese speakers (e.g. rich culture background) and fluent in English (e.g. MA degree in English translation). A one-week trial annotation phase was conducted to assess accuracy and consistency, followed by a three-week formal phase on an enterprise-level annotation platform. We ensured that none of the annotators had conflicts of interest, and their annotations were routinely cross-checked for consistency. In terms of Fleiss' Kappa, the inter-annotator agreement on Table 4 and Table 5 are 0.59 and 0.61 respectively, which is acceptable for the subjective nature of the task.

Table 4 lists the results. Our proposed metric outperforms both LaBSE and CLIP in terms of correlation with human evaluation scores across all criteria. The positive correlation coefficients for our metric indicate a strong agreement between the multi-modal alignment metric and human judgments. This suggests that our metric is more effective in capturing the key aspects of T2I generation tasks than the other two metrics. The results clearly demonstrate the superiority of our metric in assessing the quality of translated captions for the T2I generation tasks. We also investigate the impact of object-text alignment score in our metric by removing it from the ultimate score (i.e. "$-A_{O-T}$"), which is one of the key challenges in cross-cultural T2I generation. The results confirm our hypothesis: removing the object-text alignment score drastically decreases the correlation with human judgement, indicating that the alignment is essential in assessing the translated caption for cross-cultural T2I generation.

Performance on the C^3 Benchmark Table 5 lists the results of different data filtering approaches on the proposed C^3 benchmark. We also list the results of randomly sampling 300K instances for reference. Clearly, all fine-tuned models achieve significantly better performance than the vanilla model that is trained only on the English-centric data, which confirms the necessity of fine-tuning on the target cultural data for cross-cultural generation. All filtering approaches with certain metrics outperform the randomly sampling strategy, demonstrating that these metrics are reasonable for filtering low-quality instances. Our metric obtains the best results under all criteria by maintaining high-quality instances for fine-tuning. Figure 6 shows some example images generated by different models. The vanilla diffusion model fails to generate Chinese-cultural elements, which can be greatly mitigated by the fine-tuned models. While CLIP and Our models successfully generate all the objects in the captions (e.g. "tea ceremony with an expert" and "winding pathways, carefully placed rocks, and lush vegetation"), the elements in our images appear more natural and better-integrated. We attribute the strength of our approach to the explicit consideration of object-text alignment in data filtering. It is also worthy noting that the proposed C^3 benchmark can distinguish different models by identifying model-specific weaknesses.

5 Conclusion and Future Work

In this work, we build a C^3 benchmark of challenging textual prompts to generate images in Chinese cultural style for T2I models that are generally trained on the English data of Western cultural

[2] https://github.com/CompVis/stable-diffusion.
[3] https://huggingface.co/datasets/laion/laion2B-multi.

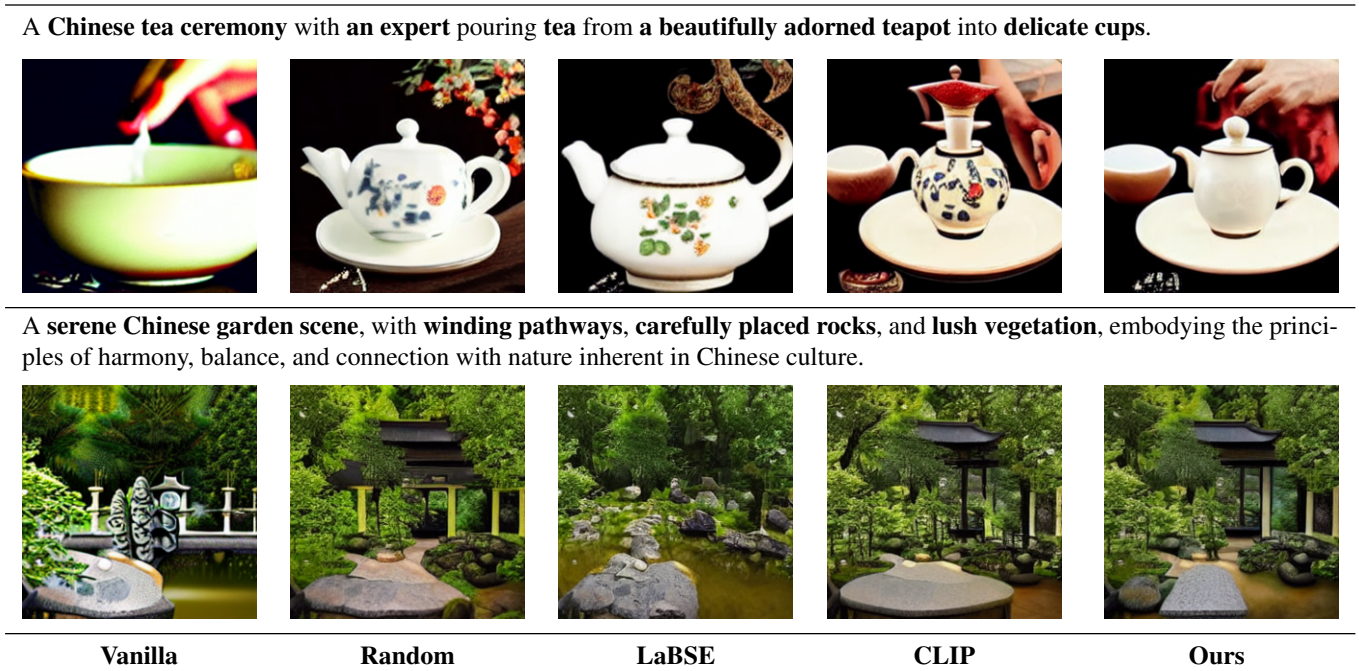


Figure 6. Example images generated by vanilla and fine-tuned diffusion models. We highlight in **bold** the objects in the caption.

elements. We demonstrate how the benchmark can be used to assess a T2I model's ability of cross-cultural generation from different perspectives, which reveal that the object generation is one of the key challenges. Based on the observation, we propose a multi-modal approach that explicitly considers object-text alignment for filtering fine-tuning data, which can significantly improves cross-cultural generation over existing metrics. Future work include extending the C^3 benchmark to more non-English cultures (e.g. Arabic culture), validating our findings with more T2I models such as DALL-E 2 [24]. Additionally, exploring the integration of our benchmark with large multi-modal models could provide further insights into how these models handle cross-cultural generation tasks, particularly in representing complex visual and textual information [19, 15, 12]. We also see potential in applying our approach to multi-modal machine translation tasks, where the accurate representation of culturally specific objects in images is critical for successful translation [31, 10, 11]. Moreover, expanding our benchmark to evaluate multi-modal question answering systems could help assess their ability to reason about culturally diverse visual content [27, 14]. Finally, assessing the perceptual capabilities of multi-modal models in understanding and generating culturally nuanced images is another promising direction for future work [33].

Limitations

This study, while providing valuable insights into the performance of T2I models in cross-cultural contexts, has several limitations that merit discussion. One notable limitation is our reliance on human annotators for the evaluation of T2I models. Although this approach offers nuanced understanding, it incurs higher costs and lacks the scalability of automated methods. Additionally, the dataset generated by GPT-4 may carry inherent language biases, particularly an English-centric perspective on cultural elements. Despite efforts to mitigate this through expert reviews, the potential for bias persists. This limitation points to the broader issue in AI research regarding the balance between automated data generation and the need for cultural neutrality and sensitivity. Moreover, our focus on Chinese culture, while grounded in our expertise, also brings to light the generalizability of our findings. The specific cultural focus may not fully translate to other cultural contexts or languages. This aspect emphasizes the delicate nature of representing and understanding cross-cultural nuances in T2I models. The definition and accurate representation of cross-culture itself present a complex challenge that our study only begins to address.

References

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[3] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, et al. PALI: A jointly-scaled multilingual language-image model. In *ICLR*, 2023.

[4] Z. Chen, G. Liu, B.-W. Zhang, F. Ye, Q. Yang, and L. Wu. Altclip: Altering the language encoder in clip for extended language capabilities, 2022.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *CILR*, 2021.

[7] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In *ACL*, 2022.

[8] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng, Y. S. Sun, L. Chen, H. Tian, H. Wu, and H. Wang. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *CVPR*, 2023.

[9] G. Huang, L. Liu, X. Wang, L. Wang, H. Li, Z. Tu, C. Huang, and S. Shi. TranSmart: A Practical Interactive Machine Translation System. *arXiv*, 2021.

[10] Z. Lan, J. Yu, X. Li, W. Zhang, J. Luan, B. Wang, D. Huang, and J. Su. Exploring better text image translation with multimodal codebook. *arXiv preprint arXiv:2305.17415*, 2023.

[11] Z. Lan, L. Niu, F. Meng, J. Zhou, M. Zhang, and J. Su. Translatotron-v (ison): An end-to-end model for in-image machine translation. *arXiv preprint arXiv:2407.02894*, 2024.

[12] H. Li, S. Li, D. Cai, L. Wang, L. Liu, T. Watanabe, Y. Yang, and S. Shi. Textbind: Multi-turn interleaved multimodal instruction-following. *arXiv preprint arXiv:2309.08637*, 2023.

[13] Y. Li, C.-Y. Chang, S. Rawls, I. Vulić, and A. Korhonen. Translation-enhanced multilingual text-to-image generation. In *ACL*, 2023.

[14] Y. Li, L. Wang, B. Hu, X. Chen, W. Zhong, C. Lyu, and M. Zhang. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*, 2023.

[15] B. Liu, C. Lyu, Z. Min, Z. Wang, J. Su, and L. Wang. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. *arXiv preprint arXiv:2312.01714*, 2023.

[16] B. Liu, L. Wang, C. Lyu, Y. Zhang, J. Su, S. Shi, and Z. Tu. On the cultural gap in text-to-image generation. *arXiv preprint arXiv:2307.02971*, 2023.

[17] Z. Liu, Y. Shin, B.-C. Okogwu, Y. Yun, L. Coleman, P. Schaldenbrand, J. Kim, and J. Oh. Towards Equitable Representation in Text-to-Image Synthesis Models with the Cross-Cultural Understanding Benchmark (CCUB) Dataset. *arXiv*, 2023.

[18] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[19] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[21] V. Prabhakaran, R. Qadri, and B. Hutchinson. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*, 2022.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[23] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[24] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[27] D. Romero, C. Lyu, H. A. Wibowo, T. Lynn, I. Hamed, A. N. Kishore, A. Mandal, A. Dragonetti, A. Abzaliev, A. L. Tonja, B. F. Balcha, C. Whitehouse, C. Salamea, D. J. Velasco, D. I. Adelani, D. L. Meur, E. Villa-Cueva, F. Koto, F. Farooqui, F. Belcavello, G. Batnasan, G. Vallejo, G. Caulfield, G. Ivetta, H. Song, H. B. Ademtew, H. Maina, H. Lovenia, I. A. Azime, J. C. B. Cruz, J. Gala, J. Geng, J.-G. Ortiz-Barajas, J. Baek, J. Dunstan, L. A. Alemany, K. R. Y. Nagasinghe, L. Benotti, L. F. D'Haro, M. Viridiano, M. Estecha-Garitagoitia, M. C. B. Cabrera, M. Rodríguez-Cantelar, M. Jouitteau, M. Mihaylov, M. F. M. Imam, M. F. Adilazuarda, M. Gochoo, M.-E. Otgonbold, N. Etori, O. Niyomugisha, P. M. Silva, P. Chitale, R. Dabre, R. Chevi, R. Zhang, R. Diandaru, S. Cahyawijaya, S. Góngora, S. Jeong, S. Purkayastha, T. Kuribayashi, T. Jayakumar, T. T. Torrent, T. Ehsan, V. Araujo, Y. Kementchedjhieva, Z. Burzo, Z. W. Lim, Z. X. Yong, O. Ignat, J. Nwatu, R. Mihalcea, T. Solorio, and A. F. Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.

[28] S. Ruan, Y. Zhang, K. Zhang, Y. Fan, F. Tang, Q. Liu, and E. Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13960–13969, 2021.

[29] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[30] M. Saxon and W. Y. Wang. Multilingual conceptual coverage in text-to-image models. In *ACL*, 2023.

[31] H. Shen, L. Shao, W. Li, Z. Lan, Z. Liu, and J. Su. A survey on multi-modal machine translation: Tasks, methods and challenges. *arXiv preprint arXiv:2405.12669*, 2024.

[32] L. Struppek, D. Hintersdorf, and K. Kersting. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*, 2022.

[33] Z. Wang, L. Wang, Z. Zhao, M. Wu, C. Lyu, H. Li, D. Cai, L. Zhou, S. Shi, and Z. Tu. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation. *arXiv preprint arXiv:2311.16511*, 2023.

[34] J. Wu, J. Wang, Z. Yang, Z. Gan, Z. Liu, J. Yuan, and L. Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.

[35] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[36] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[37] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[38] M. Zhu, P. Pan, W. Chen, and Y. Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019.