On the Cultural Gap in Text-to-Image Generation

Anonymous ACL submission

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 (Rombach et al., 2022). 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 (Prabhakaran et al., 2022; Struppek et al., 2022). 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.

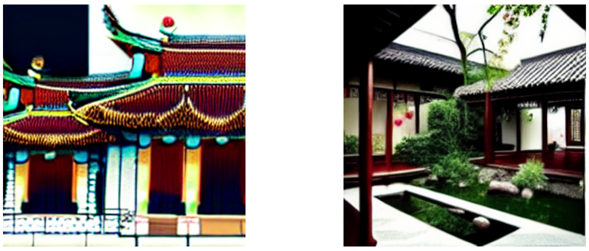


Figure 1: Comparison of the original stable diffusion (left) and the stable diffusion fine-tuned on the dataset filtered by our approach (right) for generating cross-cultural images with Chinese elements based on the prompt *A garden with typical Chinese architecture and design elements*. The example 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 (Rombach et al., 2022; Ramesh et al., 2022), 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 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) (Reed et al., 2016; Xu et al., 2018; Zhang et al., 2017), 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 (Ramesh et al., 2022, 2021; Rombach et al., 2022; Saharia et al., 2022). Previous research on text-guided image generation mainly focused on improving the understanding of complex text descriptions (Zhu et al., 2019; Ruan et al., 2021) or the quality of generated images (Saharia et al., 2022). 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. (2022) extent the text encoder of diffusion model with a pre-trained multilingual text encoder XLM-R. Li et al. (2023) mitigated the language gap by translating English captions to other languages with neural machine translation systems. Chen et al. (2023) 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 (2023) 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. (2023) 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 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

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. Figure 2 shows the benchmark details.

3.2 Evaluating Difficulty of C^3 Benchmark

To evaluate the difficulty of the C^3 benchmark, we compare with the commonly-used COCO Captions dataset (Chen et al., 2015), 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 2 shows the details of the sampled COCO

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 1: Five seed captions for constructing benchmark.

	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.

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 (Saharia et al., 2022; Feng et al., 2023). Figure 3 lists the comparison results. Clearly, 78% of the generated images on COCO are rated above average (“≥ 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 4 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”) 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.

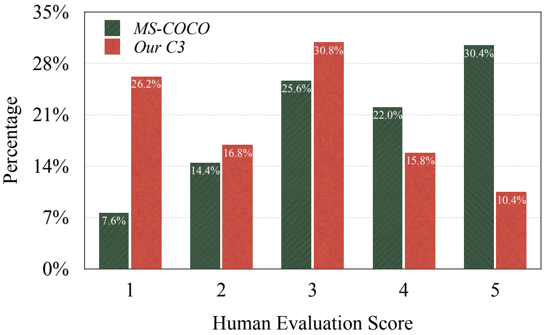
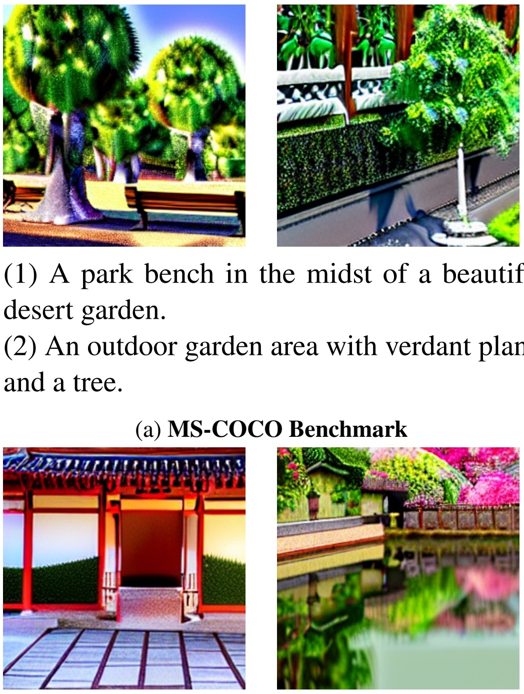


Figure 3: Human scoring results of Stable Diffusion on the widely-used MS-COCO and the proposed C^3 benchmarks.

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 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



(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 image.

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.

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

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.

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 2 lists an example of using the criteria to evaluate the image in Figure 1 (left panel). Table 5 in Appendix 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 (Lyu et al., 2023), 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 and Meteor

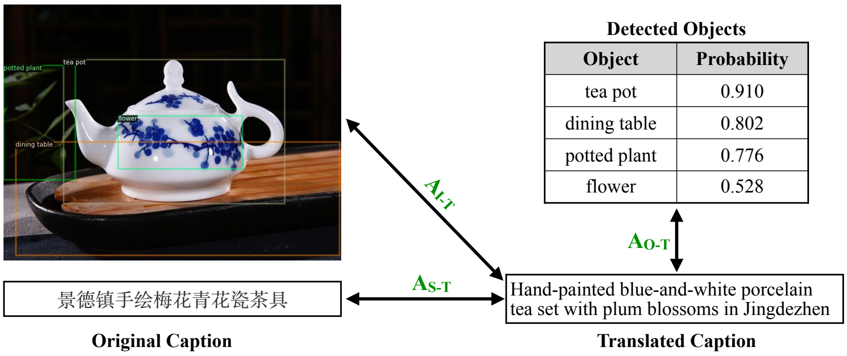


Figure 5: Framework of our filtering metric that measures the quality of the translated caption with three alignment scores: 1) A_{S-T} for aligning the original caption; 2) A_{I-T} for aligning the image; and 3) A_{O-T} for aligning the detected objects.

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 (Zhang et al., 2019), 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. (2022) 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 (Radford et al., 2021), 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 the 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 $\mathcal{E} \in \mathbb{R}^h$ (Devlin et al., 2019) to the corresponding representations:

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

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 $\mathcal{V} \in \mathbb{R}^h$ (Dosovitskiy et al., 2021) into a representation vector:

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

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:

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

Similarly, we calculate the other two alignment scores by:

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

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

The ultimate score is a combination of the above alignments:

$$A = A_{S-T} + A_{I-T} + A_{O-T} \qquad (6)$$

The score A reflects the quality of the translated captions by considering both their textual and visual

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 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.

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 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 (Wu et al., 2022) with prediction probability > 0.5.

4.3 Experiments

Experimental Setup We conduct experiments with the Stable Diffusion v1-4 model (Rombach et al., 2022).[1] For fine-tuning the diffusion model on the Chinese cultural data, we choose the Chinese subset (*laion2b-zh*) of the *laion2b-multi* dataset[2], comprising a total of 143 million image-text pairs. We translate all image captions into English using an online translation system TranSmart (Huang et al., 2021) (https://transmart.qq.com).

We filter the full *laion-zh* to 300K instances with different strategies, including 1) the text-text alignment score **LaBSE** (Feng et al., 2022); 2) the image-text alignment score **CLIP** (Radford et al., 2021); 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 (Loshchilov and Hutter, 2019) 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 Appendix 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.

Table 3 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.

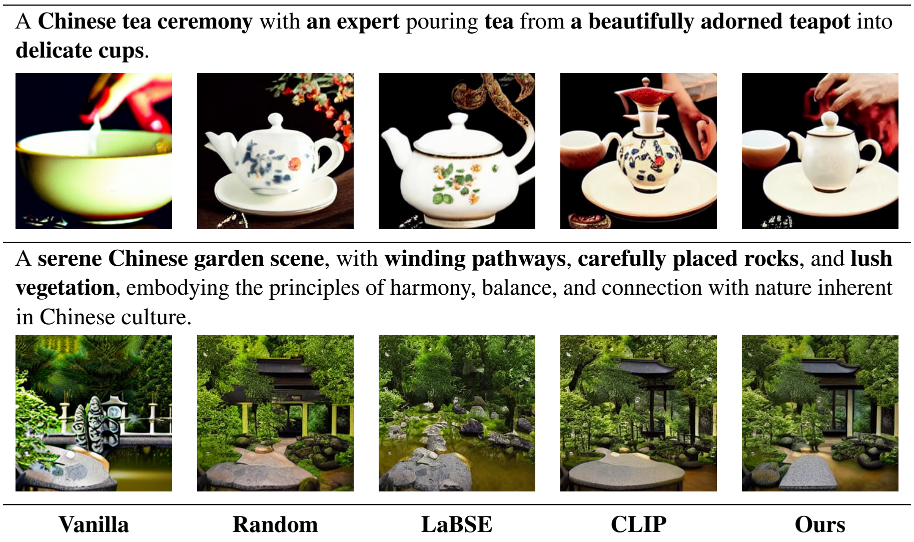
Performance on the C^3 Benchmark Table 4 lists the results of different data filtering approaches on the proposed C^3 benchmark. We also list the results of randomly sampling 300K instances for

[1] https://github.com/CompVis/stable-diffusion.

[2] https://huggingface.co/datasets/laion/laion2B-multi.

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**

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


A **Chinese tea ceremony** with **an expert** pouring **tea** from **a beautifully adorned teapot** into **delicate cups**.

A **serene Chinese garden scene**, with **winding pathways**, **carefully placed rocks**, and **lush vegetation**, embodying the principles of harmony, balance, and connection with nature inherent in Chinese culture.

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

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 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 (Ramesh et al., 2022).

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

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2023. PALI: A jointly-scaled multilingual language-image model. In *ICLR*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *ACL*.

Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun Sun, Li Chen, Hao Tian, Hua Wu, and Haifeng Wang. 2023. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *CVPR*.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. TranSmart: A Practical Interactive Machine Translation System. *arXiv*.

Yaoyiran Li, Ching-Yun Chang, Stephen Rawls, Ivan Vulić, and Anna Korhonen. 2023. Translation-enhanced multilingual text-to-image generation. In *ACL*.

Zhixuan Liu, Youeun Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. 2023. Towards Equitable Representation in Text-to-Image Synthesis Models with the Cross-Cultural Understanding Benchmark (CCUB) Dataset. *arXiv*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *arXiv preprint arXiv:2211.13069*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

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 *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.

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.

Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. 2021. 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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photo-realistic text-to-image diffusion models with deep language understanding. In *NeurIPS*.

Michael Saxon and William Yang Wang. 2023. Multilingual conceptual coverage in text-to-image models. In *ACL*.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2022. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. 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.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. 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.

6 Appendix

6.1 Human Evaluation Guidelines

Table 5 provides a detailed guideline for evaluating the generated images on the C^3 benchmark. The evaluation is based on six criteria: Object Presence, Object Localization, Cultural Appropriateness, Visual Aesthetics, Semantic Consistency, and Cohesion. Each criterion is scored on a scale of 1 to 5, with 5 being the highest score.

- **Object Presence:** This criterion assesses whether all essential objects described in the caption are present and clearly recognizable in the generated image.
- **Object Localization:** This criterion evaluates whether the spatial arrangement of objects in the image accurately represents the arrangement described in the caption.
- **Cultural Appropriateness:** This criterion measures whether the cultural style and context described in the caption are clearly and consistently reflected in the image.
- **Visual Aesthetics:** This criterion assesses the visual appeal and composition of the image, including color harmony, contrast, and image sharpness.
- **Semantic Consistency:** This criterion evaluates the consistency between the image and the caption, i.e., whether all elements in the image align with and accurately represent the text.
- **Cohesion:** This criterion measures the coherence and unity in the image, i.e., whether all elements in the image appear natural and well-integrated, creating a seamless visual scene.

Each of these criteria is crucial for evaluating the performance of T2I models, as they collectively assess the model's ability to generate images that are not only visually appealing and semantically consistent with the input text, but also culturally appropriate and coherent.

6.2 Evaluation Guidelines for Translated Captions

Table 6 provides a detailed guideline for evaluating the translated captions associated with the images. The evaluation is based on six criteria: Adequacy, Fluency, Consistency, Relevance, Context, and Cultural Appropriateness. Each criterion is scored on a scale of 1 to 5, with 5 being the highest score.

- **Adequacy:** This criterion assesses whether the translation accurately conveys the intended meaning of the original caption.
- **Fluency:** This criterion evaluates the fluency of the translation, including grammar, syntax, and vocabulary.
- **Consistency:** This criterion measures the consistency of the translations in terms of language, tone, and style.
- **Relevance:** This criterion assesses whether the translations are relevant to the image they describe, capturing the essence of the image and all important details.
- **Context:** This criterion evaluates whether the translations provide sufficient context for the reader to understand the image and the situation in which it was taken.
- **Cultural Appropriateness:** This criterion measures whether the translations are appropriate for the target audience, demonstrating an understanding of the target culture and avoiding cultural references or language that could be offensive or confusing.

These criteria provide a comprehensive framework for evaluating the quality of the translated captions associated with the images, offering insights into the strengths and weaknesses of the translation process.

Table 5: Evaluation guidelines for generated images on the C^3 benchmark.

Score	Object Presence	Object Localization	Cultural Appropriateness
5	All essential objects are present and clearly recognizable, making the image fully consistent with the caption.	All objects are placed correctly and consistently, accurately representing the spatial arrangement described in the caption.	Cultural style and context are clearly and consistently reflected, making the image an excellent representation of the intended culture.
4	Most essential objects are present and recognizable, with only minor inconsistencies or missing details.	Most objects are placed correctly with few inconsistencies, showing a good understanding of the spatial arrangement described in the caption.	Cultural style and context are mostly well-reflected, with only minor inconsistencies or missing elements.
3	Essential objects are present, but some are missing or unclear, making the image not fully consistent with the caption.	Objects are placed reasonably well, but some inconsistencies or minor errors exist in the spatial arrangement.	Some cultural style or context is reflected, but not consistently or convincingly throughout the entire image.
2	Some essential objects are present, but not clearly recognizable or only partially visible.	Some objects are placed correctly, but most are not, showing a weak understanding of spatial arrangement.	Minimal cultural style or context is reflected, with only one or two elements hinting at the intended culture.
1	No essential objects are present in the generated image.	Objects are randomly placed with no spatial arrangement, disregarding the captions.	No cultural style or context is reflected in the generated image.

Score	Visual Aesthetics	Semantic Consistency	Cohesion
5	Excellent visual appeal and composition, with perfect color harmony, contrast, and image sharpness, resulting in a visually stunning image.	Complete consistency between the image and caption, with all elements aligning and accurately representing the text.	Complete coherence and unity in the image, with all elements appearing natural and well-integrated, creating a seamless visual scene.
4	Above average visual appeal and composition, with good color harmony, contrast, and image sharpness, making the image visually pleasing.	High consistency between the image and caption, with most elements aligning and only minor inconsistencies.	High coherence and unity in the image, with almost all elements appearing natural and well-integrated, creating a cohesive visual scene.
3	Average visual appeal and composition, with acceptable color harmony, contrast, and image sharpness, but lacking any outstanding qualities.	Moderate consistency between the image and caption, with some elements aligning but not enough to provide a strong connection.	Moderate coherence and unity in the image, with most objects appearing natural and well-integrated, but some inconsistencies are present.
2	Below average visual appeal and composition, with some issues in color harmony, contrast, or image sharpness.	Minimal consistency between the image and caption, with only one or two elements connecting the image to the text.	Minimal coherence or unity in the image, with some objects appearing out of place or detached from the scene.
1	Poor visual appeal and composition, with unbalanced colors, low contrast, and lack of image sharpness.	No consistency between the generated image and the caption, making the image unrelated to the text.	No coherence or unity in the image, with objects appearing disjointed and unnatural.

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.