# From Editing to Transcreation: Adapting Image Editor for Cultural Localization

Youyuan Lin[1] Yizhou Zhang[1] Chenhui Chu[1]
[1]Kyoto University
lin.youyuan.73v@st.kyoto-u.ac.jp
yizhang@sap.ist.i.kyoto-u.ac.jp
chu@i.kyoto-u.ac.jp

## Abstract

Image transcreation target cultural localization: edits images across cultures while preserving intent. But the scarcity of paired source–target images limits task-specific training and reduces edit fidelity. We construct human-verified transcreation pairs from MaRVL for China→Japan and China→United States, each accompanied by explicit editing instructions. We fine-tune Step1X-Edit on (source image, instruction, target image) triplets and evaluate outputs with an LLM-as-a-Judge framework. The results show that fine-tuning improves cultural relevance and visual naturalness but reduces semantic consistency, motivating constraint-aware methods to mitigate semantic drift.

## 1 Introduction

Visual semantics are often culture-dependent: clothing, text, and even composition can evoke different meanings across regions. As a result, visual content that is effective in one cultural context may be confusing or inappropriate in another, creating a barrier to multicultural communication, particularly in audiovisual media and marketing (Figure 1). In natural language processing, such challenges are commonly addressed through **transcreation**, which adapts text to a target culture while preserving intent [1]. **Image transcreation** extends this idea to the visual domain, but existing pipelines remain unreliable [2].

Recent vision-language models (VLMs) encode non-trivial cultural knowledge [3, 4]. This suggests a natural division of labor: an VLM generates culture-aware edit instructions, while an image editor executes these instructions to achieve cultural localization. However, recent instruction-following editors (e.g., SDEdit [5] and InstructPix2Pix [6]) are not specifically trained for cultural transcreation. Moreover, Strengthening the editor's ability to apply culture-focused instructions requires paired data: source images aligned with edit instructions and localized target images. Such datasets are scarce.

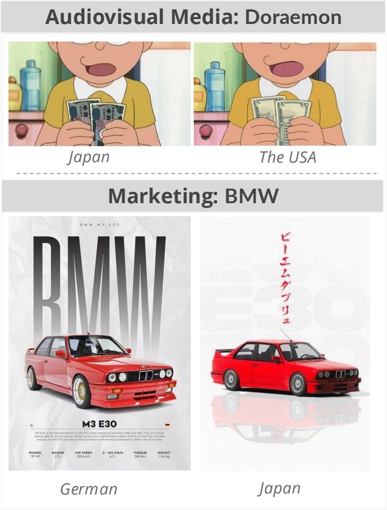


Figure 1: Transcreation cases in audiovisual media and global marketing.

To address this, we construct a human-verified synthetic transcreation dataset based on Gemini-2.5-Pro and Gemini-2.5-image-flash [7]. We then fine-tune Step1X-Edit [8] for cultural localization. Starting from 1,271 Chinese images in MaRVL [9], we build two directions (China→United States and China→Japan) and retain 543 (Zh-En) and 594 (Zh-Ja) pairs after verification. Using the resulting (source image, instruction, target image) triplets, we fine-tune Step1X-Edit and evaluate with an LLM-as-a-Judge framework. Fine-tuning improves cultural relevance and naturalness over the untuned model but reduces semantic consistency, indicating a trade-off between cultural adaptation and semantic preservation.

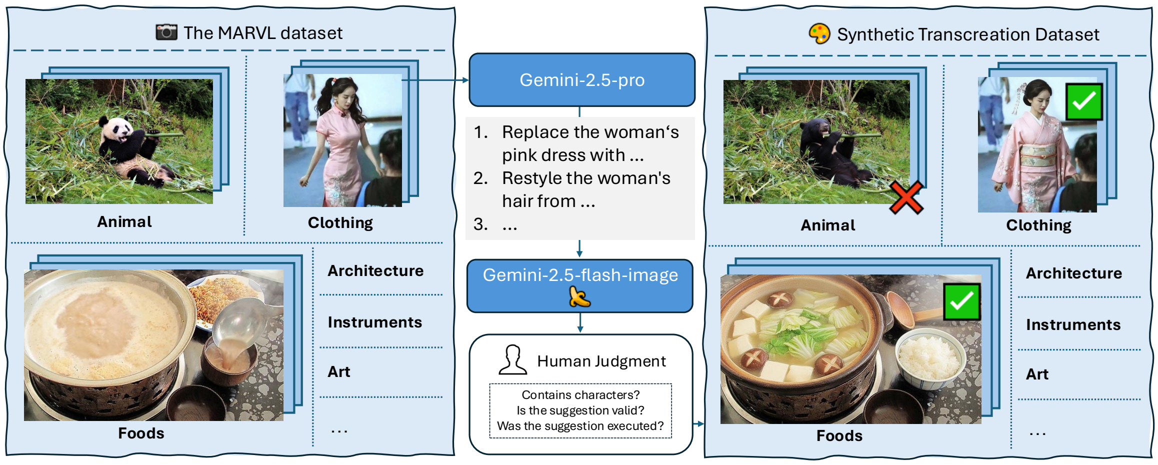


Figure 2: Dataset construction pipeline.

## 2 Dataset Collection

The overall workflow is illustrated in Figure 2. We start from MaRVL [9] for its broad coverage of culturally grounded visual concepts. We consider transcreation image from China in the Marvl dataset to Japan (Zh-Ja) and the United States (Zh-En). For each source image, Gemini-2.5-Pro generates atomic edit instructions, which Gemini-2.5-image-flash executes to produce a target-culture image [7]. Prompt details are in Appendix A.

Although Gemini-2.5-image-flash is capable of producing visually plausible results, failures remain, including semantically implausible edits, partial execution under long instruction sequences, and errors in character-level rendering. We therefore verify each example, labeling suggestion validity, execution, and whether the image contains characters. From 1,271 source images, the pipeline produces 1,089 (Zh-Ja) and 1,015 (Zh-En) edited candidates (The missing images are because of being determined not to require modification). After verification, we retain 594/543 pairs w.r.t. Zh-Ja and Zh-En directions.

We group examples into categories to support category-wise analysis, because cultural salience edits vary across visual concepts [2]. Figure 3 summarizes the category distribution over the train/val/test split. Each sample is a triplet $(I_{\text{src}}, I_{\text{target}}, c)$, where:

- $I_{\text{src}}$: Source image from the source cultural domain.
- $I_{\text{target}}$: Target image representing the desired cultural transformation.
- $c$: Natural language caption specifying the editing operations.

## 3 Experiment Setup

**Models.** We compare three settings: **Gemini** (Gemini-2.5-image-flash), **Step1X** (Step1X-Edit without fine-tuning on our dataset), and **Step1X-ft** (Step1X-Edit fine-tuned on our dataset with Eq. (1)). For all settings, edit instructions are generated by Gemini-2.5-Pro.

Our fine-tuning implementation follows the official Step1X-Edit repository.[1] Training is conducted on 2 NVIDIA RTX A6000 GPUs. Training hyperparameters are listed in Appendix B.

**Loss.** We fine-tune Step1X-Edit on the constructed triplets $(I_{\text{src}}, I_{\text{target}}, c)$ via LoRA [10] applied to the Linear layers within the Diffusion Transformer architecture. Let $z_{\text{src}}$ and $z_{\text{target}}$ denote the VAE latents of $I_{\text{src}}$ and $I_{\text{target}}$. Following the flow-matching objective used in Step1X-Edit, we sample $\epsilon \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}(0,1)$, and form the interpolation $z_t = (1-t)\epsilon + t z_{\text{target}}$. The model predicts the velocity conditioned on the reference and instruction, and we minimize:

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{(I_{\text{src}}, I_{\text{target}}, c)\\ t, \epsilon}} \left[ \left\| v_\theta(z_t, t\, ;\, z_{\text{src}}, c) - (z_{\text{target}} - \epsilon) \right\|_2^2 \right]. \tag{1}$$

**Evaluation Metrics.** We follow an LLM-as-a-Judge evaluation framework for image transcreation [11] and use Qwen-VL-8B to score each pair on a 1–5 Likert scale across:

- **Visual Change**: degree of visual modification.
- **Cultural Relevance**: alignment with target culture.
- **Semantic Consistency**: preservation of information.
- **Naturalness**: visual plausibility.

[1] https://github.com/stepfun-ai/Step1X-Edit

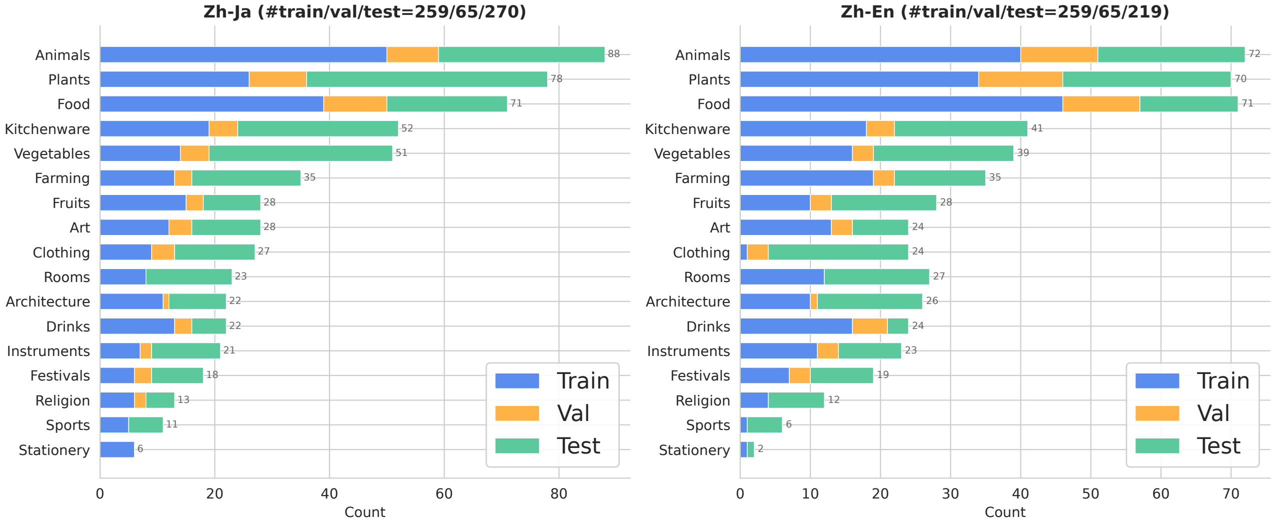
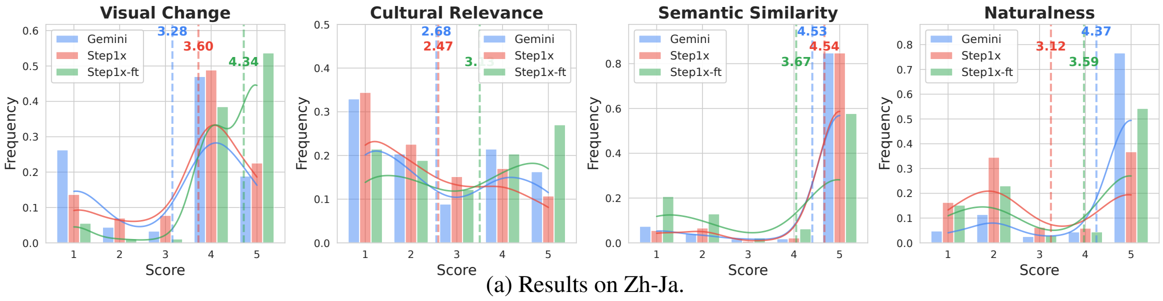
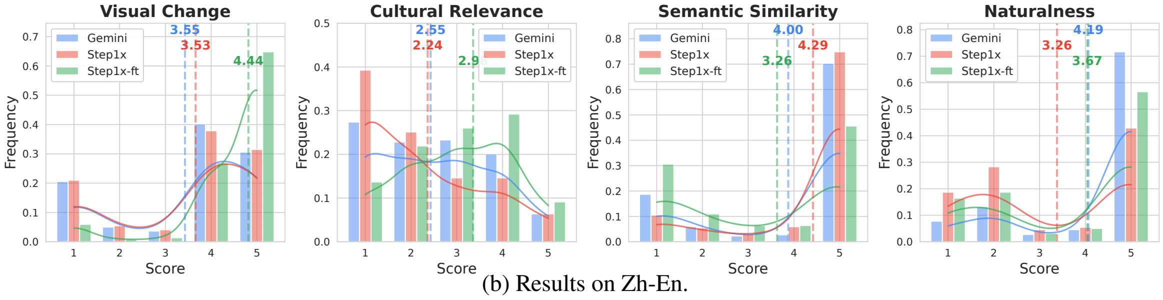


Figure 3: Dataset statistics.



(a) Results on Zh-Ja.



(b) Results on Zh-En.

Figure 4: Score distributions.

# 4 Results

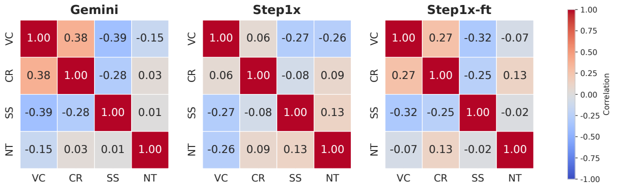
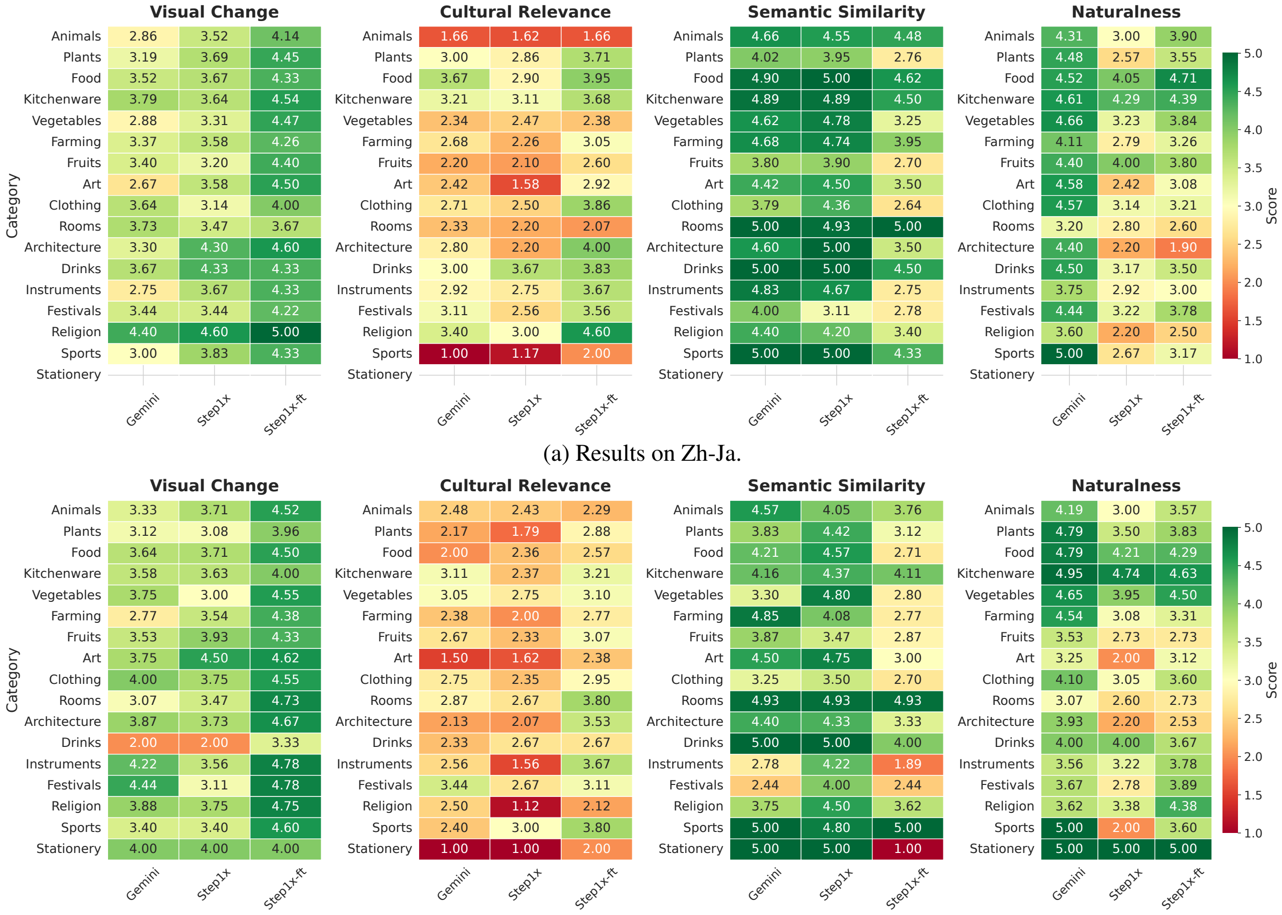


Figure 5: Correlation matrix of the four evaluation metrics aggregated across two culture pairs. VC: Visual Change, CR: Cultural Relevance, SS: Semantic Consistency, NT: Naturalness.

**Overall trends.** Figure 4 summarizes overall trends. Step1X shows a moderate cultural relevance improvement, and fine-tuning it induces more aggressive visual modifications, leading to improved cultural relevance and naturalness, but reduced semantic consistency. Gemini remains the most natural. Case studies are presented in Appendix C.

**Metric interactions.** We further draw the correlation matrix in Figure 5. The correlation analysis suggests that fine-tuning makes visual change more strongly aligned with cultural relevance, while mitigating the trade-off between visual change and naturalness, i.e., larger edits no longer systematically lead to less natural outputs. However, semantic consistency remains negatively associated with both visual change and cultural relevance, indicating that preserving meaning while achieving target-culture fit



(a) Results on Zh-Ja.

(b) Results on Zh-En.

Figure 6: Average scores by super-category.

is still an open challenge.

**Where fine-tuning helps.** Figure 6 breaks down results by super-category. Fine-tuning yields the largest gains in culturally saturated categories (e.g., Architecture, Clothing, Art, and Religion), where distinctive styles and symbols provide clear editing targets. In contrast, categories such as Animals show limited improvements in CR, likely because simple substitutions (e.g., changing bird species) are weak cultural signals and are not reliably recognized as culture-specific by the judge. These findings highlight that localization success depends not only on the editor, but also on whether the visual concept admits clear, culture-grounded realizations.

**Limitations.** Gemini edits are more natural, whereas Step1x-ft prioritizes cultural relevance via aggressive changes that risk over-editing. Furthermore, as our study focused on character-free images, our findings highlight improvements in visual elements (e.g., architecture) rather than text replacement, which remains a key challenge.

## 5 Conclusion

We constructed a paired image transcreation dataset from MaRVL Chinese subset with two target cultures (Japan and United States) with human verification. We evaluated whether Step1X can be adapted for image transcreation. We find that LoRA fine-tuning increases edit magnitude and improves cultural relevance, while partially recovering naturalness compared to the untuned model. However, these gains come with a notable reduction in semantic consistency, indicating that current editors still struggle to localize culturally salient attributes without changing category-defining content.

Future work could focus on better balance cultural adaptation and meaning preservation, as well as improved handling of text and symbols. Complementing LLM-based evaluation with targeted human studies would also help for diagnosing failure cases that automatic metrics may miss.

# Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23K28144.

# References

[1] Daniel Pedersen. Exploring the concept of transcreation–transcreation as “more than translation”. **Cultus: The Journal of intercultural mediation and communication**, Vol. 7, No. 1, pp. 57–71, 2014.

[2] Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 10258–10279, 2024.

[3] Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. Benchmarking vision language models for cultural understanding. **arXiv preprint arXiv:2407.10920**, 2024.

[4] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, et al. All languages matter: Evaluating lmms on culturally diverse 100 languages. In **Proceedings of the Computer Vision and Pattern Recognition Conference**, pp. 19565–19575, 2025.

[5] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. **arXiv preprint arXiv:2108.01073**, 2021.

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. **arXiv preprint arXiv:2211.09800**, 2022.

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. **arXiv preprint arXiv:2507.06261**, 2025.

[8] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. **arXiv preprint arXiv:2504.17761**, 2025.

[9] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. **arXiv preprint arXiv:2109.13238**, 2021.

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. **ICLR**, Vol. 1, No. 2, p. 3, 2022.

[11] Simran Khanuja, Vivek Iyer, Xiaoyu He, and Graham Neubig. Towards automatic evaluation for image transcreation. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 7034–7047, 2025.

# A Dataset Collection Prompts

We present the prompts used for collecting dataset. Variables in `{braces}` are replaced with task-specific values.

Prompt generated by following template are fed to Gemini-2.5-pro:

You are a culturally-aware image analyst for Image Transcreation.
Goal: transform imagery from {source_country} to {target_country} while preserving scene semantics.

TASK:
1) Describe the visible content concisely, focusing on culturally salient elements (attire, food, signage/text, props, symbols, architecture, scenery, color motifs).
2) Output concrete, **model-friendly** edit suggestions from {source_country} → {target_country}.
- Each suggestion MUST be an **atomic** command for a single element. If there are multiple changes, split them into separate suggestions.
- Write in natural, imperative language. Prefer starting with a clear action verb (e.g., add, remove, replace, recolor, restyle, translate_text, replace_text, move, scale, crop, blur, sharpen, brighten, darken, increase_contrast, decrease_contrast, but synonyms or short natural phrases are acceptable if clearer (e.g., "make the banner red", "change the sign to {target_country}"). Keep the action near the beginning.
- Be detailed, concise, imperative, and deterministic. No rationale, no hedging words.
- If objects are ambiguous, suggest safe neutral variants.
3) IMPORTANT: Output MUST be valid JSON conforming to the response schema.

CONSTRAINTS:
- Keep people identity and layout unless a suggestion explicitly changes them.
- Preserve lighting, pose, and composition when possible.
- Don't fabricate brands/real trademarks.

Prompt generated by following template are fed to Gemini-2.5-flash-image. `{step}` here is replaced with the suggestions generated by Gemini-2.5-pro:

Transform this image from {source_country} to {target_country} by applying ALL edits below.
Requirements:
- Apply ONLY these edits. Keep subject identity, camera angle, lighting, and composition.
- High fidelity. No extra text unless requested.
- If edits imply localization of signage/text, render clean native-language typography.
- Maintain output size close to the input.
- If an edit is ambiguous, choose a culturally neutral, widely-recognized option.

EDITS TO APPLY (execute in order):
{steps}

# B Hyperparameter Summary

We fine-tune Step1X-Edit for 32 epochs using AdamW with a learning rate of $1 \times 10^{-4}$ in bfloat16 precision. Training uses a batch size of 4 per GPU on two GPUs (effective batch size 8). LoRA is configured with rank $r = 64$ and scaling $\alpha = 32$.

# C Qualitative Examples

Figure 7 shows qualitative comparisons across Gemini, Step1X (base), and Step1X-ft on both culture pairs.

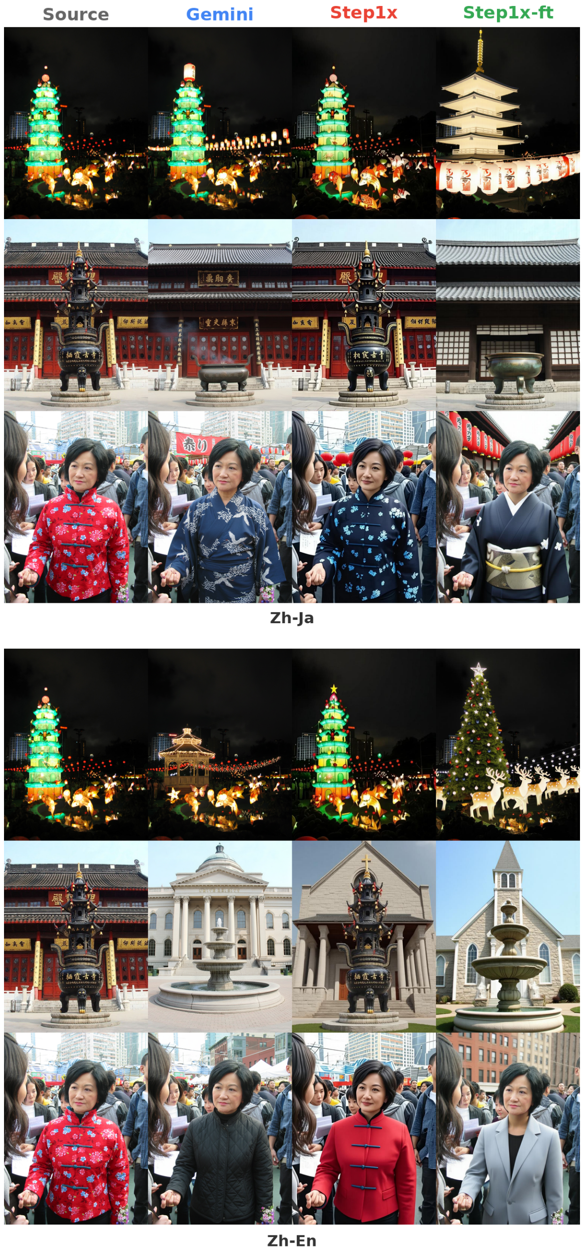


Figure 7: Qualitative examples.