Where Culture Fades:
Revealing the Cultural Gap in Text-to-Image Generation

Supplementary Material

Figure 1. **Examples of the geographical and cultural composition of the CultureBench dataset.**

The appendices provide additional details that support and extend the main paper. Appendix A describes the dataset's geographic composition, evaluation protocol, and explains the selection of specific countries. Appendix B further validates the conjecture presented in the main text. Appendix C presents further experimental results and ablation studies. Appendix D details the user study's design and conduct. Appendix E provides additional visualization results for PEA-Diffusion [3] and AltDiffusion [6] under both methods. Appendix F addresses common issues. Appendix G covers the limitations of our work. Appendix H reflects on the ethical considerations of this research.

A. Details of CultureBench Dataset

This study presents the dataset's geographical composition within a broad cultural classification framework. It encompasses major cultural spheres such as the Arab world, East Asia, continental Europe, Latin America, and parts of Africa. This approach reflects the macro-sociological context of the data sources. It does not represent an essentialist or homogenized understanding of cultures. Cultural spheres are delineated by principal nations and languages, including Arabic, Chinese, Japanese, Korean, Thai, German, Russian, Italian, Dutch, Polish, Turkish, Ukrainian, Spanish, Portuguese, and French in parts of Africa. Figure 1 illustrates an exemplar composition of this study's dataset.

Details of Data Collection. During data collection, we used Google Image Search [1] and complementary web resources. These included Wikipedia [2] and other public search platforms. We obtained publicly available images, setting keywords based on various cultural contexts to cover diverse scenarios and object types. After automated extraction, images undergo preliminary screening. This eliminates low-quality, semantically irrelevant, or copyright-infringing samples. A cross-cultural expert team then manually reviews images, prioritizing the exclusion of those conveying stereotypes or biases. In addition, these experts define and validate expert-curated cultural subdivisions within each broad cultural group. These include regional styles, ethnic or indigenous traditions, and locally distinctive visual elements. The team uses these subdivisions to guide manual filtering and annotation. This ensures an accurate representation of the intended cultural context without incorporating visual elements from other cultural backgrounds. As a result, CultureBench captures both macro-level cultural identity and finer-grained intra-cultural diversity. Through this process, we have enhanced the cultural accuracy and fairness of our cross-cultural visual data while ensuring its diversity.

B. Further Evidence on the Culture Gap

To test whether the cultural gap arises from under-activation rather than knowledge absence, we conduct a unified controlled experiment (Figure. 2). Using a culture-neutral English prompt and fixing all stochastic factors (e.g., seed and sampler), we generate images while selectively activating different culture-neuron sets identified in previous analyses. Despite identical prompts, activating China, Japan, Italy, Germany, or Ukraine neuron sets yields clearly distinct cultural styles, whereas the baseline remains culturally neutral. This controlled intervention demonstrates that the model already encodes rich culture-specific representations, and that the failure stems primarily from insufficient activation during generation rather than missing cultural knowledge.

[1] https://images.google.com
[2] https://www.wikipedia.org

Figure 2. **Further evidence on the culture gap.** For a fixed English prompt and identical sampling settings, activating different culture-neuron sets steers both PEA-Diffusion (top) and AltDiffusion (bottom) toward distinct cultural styles.

Table 1. **Comparison of accuracy between CultureVQA and human experts on the CultureBench dataset.**

Model / Group	Accuracy
Human Experts (Avg.)	94.18%
CultureVQA (Ours)	91.57%

C. More Experiments

C.1. Reliability of CultureVQA

To assess the consistency of CultureVQA with human subjective cognition, we invited 30 domain experts to participate in comparative experiments using the CultureBench dataset. Each question contained four real-world images, and only one matched the specified cultural element. CultureVQA selected one image per question, and the experts made their choices under the same conditions. We then calculated and compared the accuracy rates for both CultureVQA and the experts. As shown in Table 1, CultureVQA achieved an accuracy of 91.57%, while human experts reached an average accuracy of 94.18%, resulting in a gap of only 2.61 percentage points. This relatively small difference indicates that CultureVQA's performance closely approaches that of human experts, demonstrating high consistency with human cultural recognition in this task.

C.2. Cultural Probing Universal Type

For clarity in the main text, only the detection results from PEA-Diffusion are presented. This section additionally demonstrates the culture-sensitive layers and neurons associated with AltDiffusion for comparison purposes.

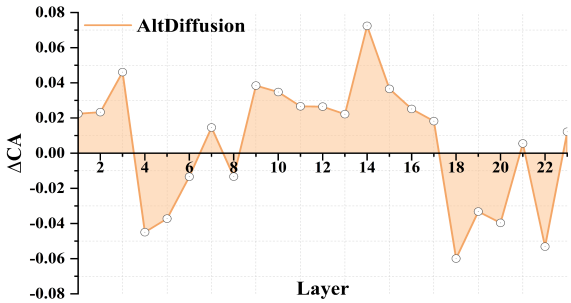

Figure 3. **AltDiffusion cultural sensitivity.** ΔCA peaks layer 14. Therefore, layer 14 is culturally sensitive.

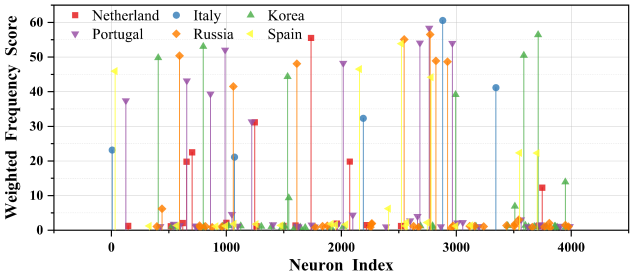

Figure 4. **AltDiffusion neuronal detection result.** The weighted frequency scores show only a few salient peaks per culture, indicating culture-specific neurons.

Culture Layer Detection. To test the generality of our detection method beyond PEA-Diffusion, we also analyze AltDiffusion. Following the main paper's probing procedure, we create paired prompts ("culture-style modifier + noun" and "noun-only"), annotate modifier and noun token groups, and extract cross-attention maps from all layers. We then compare aggregated attention from cultural modifiers to their paired nouns, yielding layerwise cultural-attention

Table 2. **Validating Culture-Sensitive Neuron Detection in AltDiffusion.** Neuronal accuracy on the test subset with "cultural style modifier + noun" prompts.

Method	CultureVQA ↑
AltDiffusion [6]	44.54
+ Masked Top-K Neurons	12.04 (-32.50)
+ Masked Random Neurons	42.45 (-2.09)

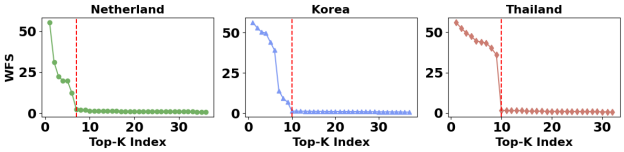

Figure 5. **Selection of Threshold K.** The red dashed vertical line indicates the chosen threshold K. Neurons to the left of the line correspond to the selected Top-K culture-sensitive neurons, while the scores to the right fall below the cutoff and are discarded.

contrast curves for AltDiffusion (Figure 3).

Culture Neuron Detection. We employ the same methodology as in the main text to localize culturally sensitive neurons within the AltDiffusion culture layer. As illustrated in Figure 4, we present culturally sensitive neurons across six distinct cultural contexts. This demonstrates our approach's robust capability to detect culturally sensitive neurons across varied architectural frameworks.

C.3. Cultural Validation in AltDiffusion

To analyze the accuracy of culture neurons in AltDiffusion, we validate these neurons for each cultural group. For all 15 cultural groups, we report CultureVQA scores in three controlled settings: (1) Baseline (no masking), (2) Masked Top-K Neurons (masking the Top-K identified culture-sensitive neurons), and (3) Random Mask (masking the same number of neurons at random). Table 2 shows that masking the Top-K neurons causes a substantial drop in CultureVQA performance. The mean score falls by 32.50 points compared to the AltDiffusion baseline. In contrast, randomly masking the same number of neurons reduces the mean score by just 2.09 points. This negligible drop stays close to the baseline. These results indicate that the culture-sensitive neurons we identify in AltDiffusion concentrate cultural information and capture cultural representations.

C.4. Selection of Threshold K

To clarify how we choose the threshold K, we visualize the weighted frequency scores (WFS) of candidate neurons and inspect their response patterns. The main factor guiding the choice of K is the presence of a small subset of neurons that exhibit prominently higher responses than the rest. As shown in Figure 5, we plot the WFS curves for three different cultural groups; in each case, a few leading neurons form clear peaks, followed by a rapid decay. We set K at the elbow before this sharp drop, so that neurons with salient high responses are retained, while the long tail of weakly responsive neurons is discarded.

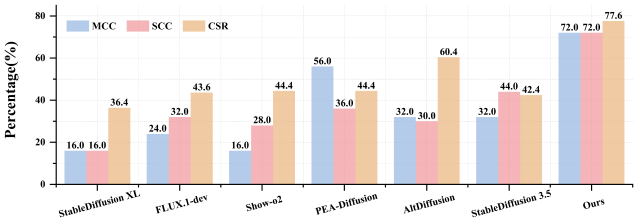

Figure 6. **User Study.** Evaluated using MCC, SCC, and CSR metrics, where higher scores indicate greater perceived realism and user preference.

C.5. Details of SAE

we set the sparsity coefficient to $\alpha = \frac{1}{32}$. We adopt a Top-K SAE with a hidden layer dimension of 4096. The model is optimized with AdamW using a learning rate of 0.0004 and a constant learning-rate schedule without warmup. We use an MSE reconstruction loss to encourage faithful feature reconstruction.

C.6. Cross-Domain Generalization Results

We evaluated our method's generalization using quantitative and qualitative tests on cue words not present in the CultureBench dataset. We used 100 out-of-domain captions to ensure the evaluation reflects true and reliable out-of-distribution performance.

C.7. User Study

Building on the quantitative and qualitative results, we conducted a human-centered user study on the CultureBench platform with 50 experts in cultural studies. Participants evaluated cultural perception using three metrics: MCC (Multi-Choice Culture), SCC (Single-Choice Culture), and CSR (Cultural Semantic Relevance, 1–5 scale). Higher scores indicate stronger cultural alignment and preference. As shown in Figure 6, our method outperforms all baselines across all three metrics. Notably, the human-rated CSR score reaches 77.6, significantly higher than the second-best score of 60.4, highlighting a clear advantage in cultural semantic fidelity. This human-evaluated CSR metric further strengthens our evaluation pipeline. Overall, these results confirm that our method can generate culturally aligned content accurately and consistently.

Quantitative Results. Table 3 shows that our method consistently improves cross-domain cultural accuracy for both PEA-Diffusion and AltDiffusion. Under the zero-training setting, our approach yields gains of +7.00 and +5.00 CultureVQA points over PEA-Diffusion and AltDiffusion, re-

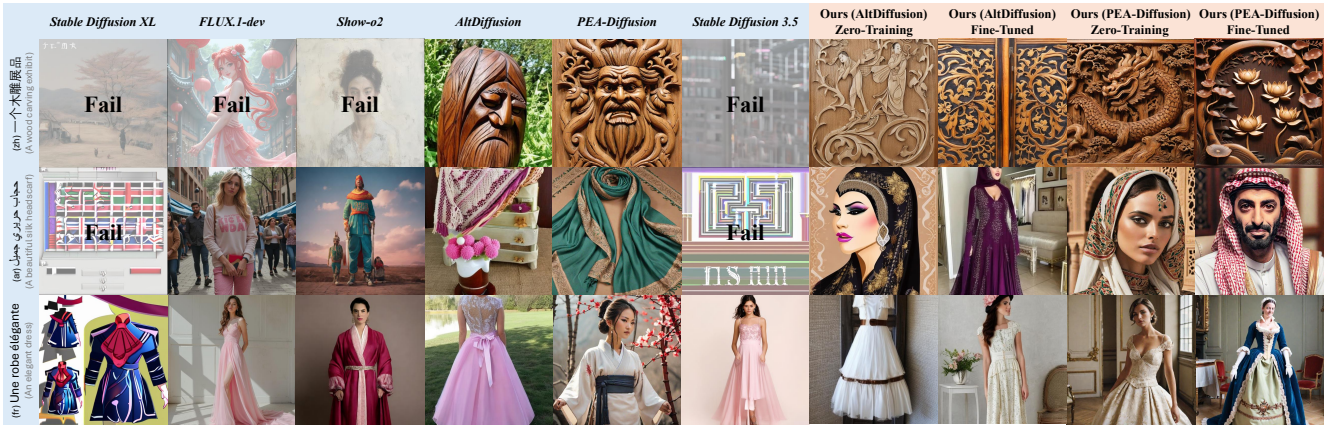

Figure 7. **Cross-domain qualitative experiments.** Our approach generates images that are more culturally appropriate.

Table 3. **Quantitative analysis across domains.** Generate using 100 captions from outside the domain and calculate the performance of CultureVQA. **Bold** rows mark the highest performance within each baseline group.

Method	CultureVQA ↑
StableDiffusion XL [4]	10.00
FLUX.1-dev [2]	17.00
Show-o2 [5]	15.00
PEA-Diffusion [3]	15.00
AltDiffusion [6]	15.00
StableDiffusion 3.5 [1]	17.00
Baseline: PEA-Diffusion	
Ours (Zero-Training)	22.00 (+7.00)
Ours (Fine-Tuned)	**35.00** (+20.00)
Baseline: AltDiffusion	
Ours (Zero-Training)	20.00 (+5.00)
Ours (Fine-Tuned)	**32.00** (+17.00)

spectively, outperforming all existing off-the-shelf models. When fine-tuned, the improvements increase to +20.00 and +17.00 points for PEA-Diffusion and AltDiffusion, respectively, with our method achieving the highest accuracy within each baseline group. These results show that the culture-sensitive neurons identified by our method provide a substantial, transferable cultural prior, enabling robust generalization to captions outside the training distribution.

Qualitative Results. As illustrated in Figure 7, when prompted with the Chinese instruction "*A wood carving exhibit*", most models exhibited semantic comprehension errors. Their outputs were inconsistent with the prompt's meaning. Our approach, however, maintained both semantic and cultural coherence. In contrast to Arabic instructions like "*A beautiful silk headscarf*" or French prompts such as "*An elegant dress*", our approach offers greater precision. It captures key textual details regarding fabric texture, style, and aesthetic quality. It also adeptly integrates regional cultural symbols and aesthetic preferences in color coordination, pattern design, and figure representation. This enables the generation of images better suited to local cultural contexts. Taken together, these qualitative findings indicate that our approach is not only more robust and expressive but also more culturally sensitive across languages and regions. It substantially improves cross-linguistic and cross-cultural text-image alignment and image generation, thereby further validating its effectiveness and broad applicability in real-world, culturally diverse scenarios.

D. Detail of User Study

As shown in Figure 8, we conducted a controlled user study to evaluate the perceptual effectiveness of this method. Fifty participants, aged 25-47 (23 female), completed three sequential perceptual tasks. Each task assessed the subjective performance of generated images, focusing on cultural alignment and preference.

Multi-Choice Culture. During the Multi-Choice Culture (MCC) mission, we first used the noun-only prompt within CultureBench and fed it to multiple image generation models, obtaining numerous candidate images. Next, participants were asked to select all images they deemed representative of a given target culture. This selection process then measured each model's capacity to generate culturally coherent images without textual cultural cues.

Single-Choice Culture. Single-Choice Culture (SCC) defines cultural perception as a forced-choice task of cultural identification. For each noun-only prompt, we generate an image using a specific model. Participants view the model-generated image and are asked to infer its most probable cultural category. We provide a fixed set of 16 options: 15 predefined cultural categories plus an additional 'Uncertain' option. Participants must select a single option based precisely on the image's visual content.

Cultural Semantic Relevance. Cultural Semantic Rele-

Figure 8. **Example interface of the user study comprising three task types.** MCC (top), SCC (bottom-left), and CSR (bottom-right).

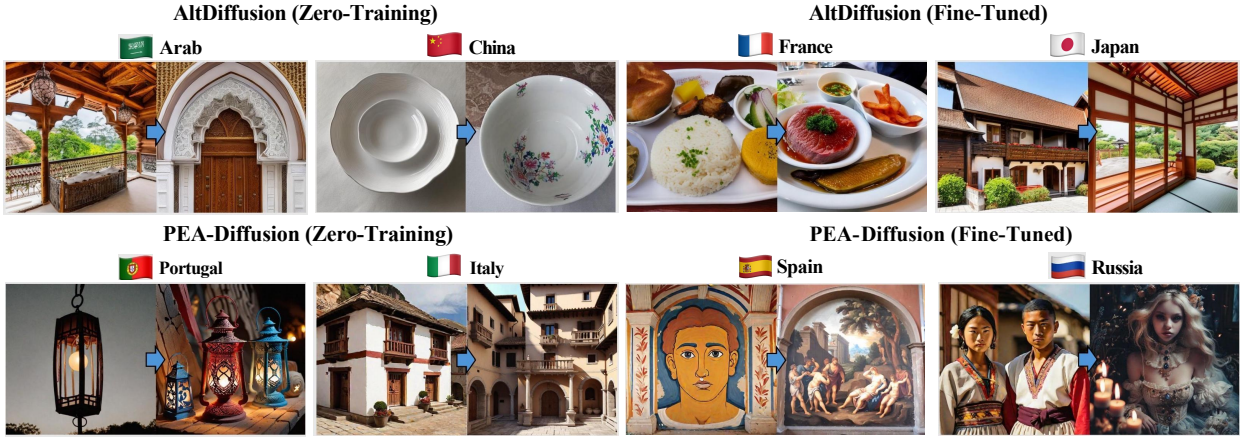

Figure 9. **More Results.** Further examples of results generated by AltDiffusion and PEA-Diffusion under different methods.

vance (CSR) is employed to evaluate the subjective cultural semantic alignment of images within a given target culture. For each CSR item, we present an image generated by a specific model, explicitly informing participants of its stylistic origin. Participants are required to rate the image's alignment with the target culture within the context of that model's style, on a scale of 1 to 5.

E. More Results

To demonstrate the effectiveness of our approach more intuitively, we generated both an unenhanced baseline image and an enhanced image using the same prompt, while keeping the model's seed and other generative parameters fixed. The comparison between the two is shown in Figure 9. The results show that the enhanced images more faithfully reflect the cultural context, including scene elements, character depictions, and fine textures. This leads to more accurate visual expressions of the target culture and further highlights the effectiveness of our approach in achieving cultural consistency.

F. More Discussion

▷ ***Q1. What do we mean by "cultural consistency" in this work?*** We define cultural consistency as the degree to which an image from a multilingual prompt shows statistically grounded, contextually appropriate cultural cues tied to the target language's socio-cultural context, rather than merely literal semantic correctness. Our approach uses (i) a fixed language–region mapping, (ii) observable, moderate visual grounding (such as architecture, clothing, artifacts), and (iii) expert screening to remove stereotypical or inappropriate cues. These elements turn a vague idea into an operational objective that can be measured and optimized.

▷ ***Q2. Why is CultureBench designed as a medium-scale diagnostic benchmark rather than a larger dataset?*** CultureBench is deliberately scaled to a moderate size, as it serves as a controlled diagnostic benchmark for assessing the cultural behaviour of multilingual text-to-image mod-

els rather than functioning as a pre-training corpus: Approximately 7.9k images, meticulously annotated across 15 language-culture regions, suffice to reveal systemic cultural omissions under "noun-only" prompts and support neural-level analysis, whilst ensuring the feasibility of high-quality, de-stereotyped annotation. We thus regard it as an extensible starting point rather than a comprehensive catalogue covering all cultures.

▷ ***Q3. Why do we rely on CultureVQA, and how reliable is it as an evaluation metric?*** We adopt CultureVQA because it directly addresses the challenge of evaluating whether images accurately reflect the nuanced cultural attributes encoded in multilingual prompts. Unlike existing automatic metrics (e.g., FID, CLIPScore), which focus on pixel-level similarity or general correspondence, CultureVQA frames evaluation as a semantic recognition problem, allowing assessment of cultural correctness. By using a VQA-style cultural identification task, CultureVQA enables consistent, scalable, and multi-class evaluation across 15 cultural regions. Its reliability is supported by a human–model consistency study (Appendix C.1).

▷ ***Q4. What evidence supports the hypothesis that cultural knowledge exists but is under-activated?*** Our claim is posed as a hypothesis and supported by converging empirical evidence rather than a formal proof. First, explicit culture-style modifiers (e.g., "*Italian architecture*", "*a person in traditional Chinese clothing*") consistently trigger culturally grounded generations. In contrast, noun-only prompts collapse to neutral, culture-agnostic prototypes, indicating that the cultural capability is present but insufficiently activated. Second, masking the Top-K neurons identified by our probing method results in a sharp drop in CultureVQA performance, whereas masking the same number of random neurons has minimal effect, suggesting a causal relationship between these neurons and cultural semantics. Third, attention-based probing reveals a stable culture-sensitive layer in which "culture-style modifier + noun" and "noun-only" prompts yield distinct activation patterns. Together, these findings show that the model has cultural representations, but they are under-activated without explicit cues.

▷ ***Q5. Why are our neuron- and layer-level interventions reasonable, and do they over-claim causality?*** Our neuron- and layer-level interventions are deliberately minimal and localized: both the zero-training neuron amplifier and the fine-tuned layer enhancer operate only within the culture-sensitive layer and within a small subspace identified by our probing method, without modifying the diffusion backbone. Adjusting these neurons reliably improves CultureVQA scores and human-perceived cultural fidelity, while leaving CLIPScore, ImageReward, LPIPS, and visual diversity essentially unchanged. This indicates that the intervention is targeted rather than disruptive. Importantly, we do not claim to establish definitive causal attributions for individual neurons. Instead, we frame our approach as a lightweight and interpretable control mechanism that leverages empirically responsive subspaces. It offers practical gains in cultural consistency while serving as a useful starting point for future, more formal causal analyses.

▷ ***Q6. Is testing only on CultureBench lacking external validation?*** We agree that relying solely on CultureBench would limit external validation, which is why we include a cross-domain experiment using 100 out-of-distribution captions not appearing in CultureBench (Appendix C.6). Our method still yields consistent gains in CultureVQA under both zero-training and fine-tuning in this out-of-domain setting, showing that the improvement is not tied to CultureBench's specific prompts.

▷ ***Q7. This paper uses ChatGPT data to refine cultural products, but could bias and errors be introduced?*** The use of ChatGPT-generated text may introduce biases; hence, we have implemented multiple dedicated measures in our data construction process to minimise these risks. Firstly, the ChatGPT content we utilise is strictly confined to "culture-style modifier + noun". All corresponding images are sourced exclusively from publicly available real-world reference materials and undergo multi-stage human expert review to eliminate stereotypes, semantic errors, and culturally inaccurate cues. Furthermore, the ChatGPT-generated modifiers themselves undergo expert filtering to ensure they do not contain stereotypical or inappropriate cultural descriptions.

▷ ***Q8. When switching generative models, is it necessary to re-execute the "Cultural Sensitivity Layer and Cultural Neuron Detection" step within our methodology?*** Whether re-detection is required depends primarily on whether the text encoder of the new model has changed. If the new generative model continues to use the same text encoder as in our experiments, re-detection is generally unnecessary; if the text encoder has been altered, the detection steps must be re-executed.

▷ ***Q9. Mean square error (MSE) cannot assess cultural consistency, so is it appropriate to use it?*** MSE is not in itself an effective metric for measuring cultural consistency, and we therefore did not employ it as an evaluation criterion. Within our framework, MSE serves as the training signal for specific layer enhancers, rather than the objective we use to gauge cultural consistency.

▷ ***Q10. What are the roles of ϵ and β?*** ϵ is set to 0 by default, since activations greater than zero are treated as neuron firing. β is a small stability constant added to the denominator to prevent it from becoming zero when a neuron has zero or very few activations.

G. Limitation

CultureBench currently covers 15 cultural regions, but this still reflects only a limited portion of global cultural diversity. Because the benchmark is built from publicly available image resources, some cultures, especially those from low-resource or marginalized communities, remain underrepresented. We emphasize that the currently included cultural regions are carefully curated and remain valid, but they do not yet exhaust the diversity within each region or across countries. In future iterations, we intend to identify underrepresented cultural groups by partnering with relevant organizations, actively seeking additional image sources, and introducing finer intra-cultural subdivisions, thereby expanding the benchmark to more countries and territories for a more comprehensive and inclusive evaluation suite.

H. Ethics Statement

In this research, we acknowledge the potential misuse of image synthesis techniques, such as ours, for generating deceptive content and spreading disinformation, a serious concern we address explicitly. However, we also note the substantial progress made in detection and prevention mechanisms in this domain. Our framework supports critical research initiatives and encourages third-party oversight, aiming to strike a balance between technological advancement and security considerations. This balanced approach promotes responsible deployment while preserving the innovation potential.

References

[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 4

[2] Black Forest Labs. black-forest-labs/flux github page, 2024. 4

[3] Jian Ma, Chen Chen, Qingsong Xie, and Haonan Lu. Peadiffusion: Parameter-efficient adapter with knowledge distillation in non-english text-to-image generation. In *European Conference on Computer Vision*, pages 89–105. Springer, 2024. 1, 4

[4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 4

[5] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025. 4

[6] Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. Altdiffusion: A multilingual text-to-image diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6648–6656, 2024. 1, 3, 4