Decomposing Subject-Driven Image Generation via Intermediate Structural Prediction

Supplementary Material

Hanzhong Guo Yizhou Yu

School of Computing and Data Science, The University of Hong Kong

hanzhong@connect.hku.hk, yizhouy@acm.org

A. Additional Qualitative Results

This section provides additional qualitative results to further demonstrate the effectiveness of our method. We first compare our approach against existing baseline models and then showcase more generation results for objects both with and without text.

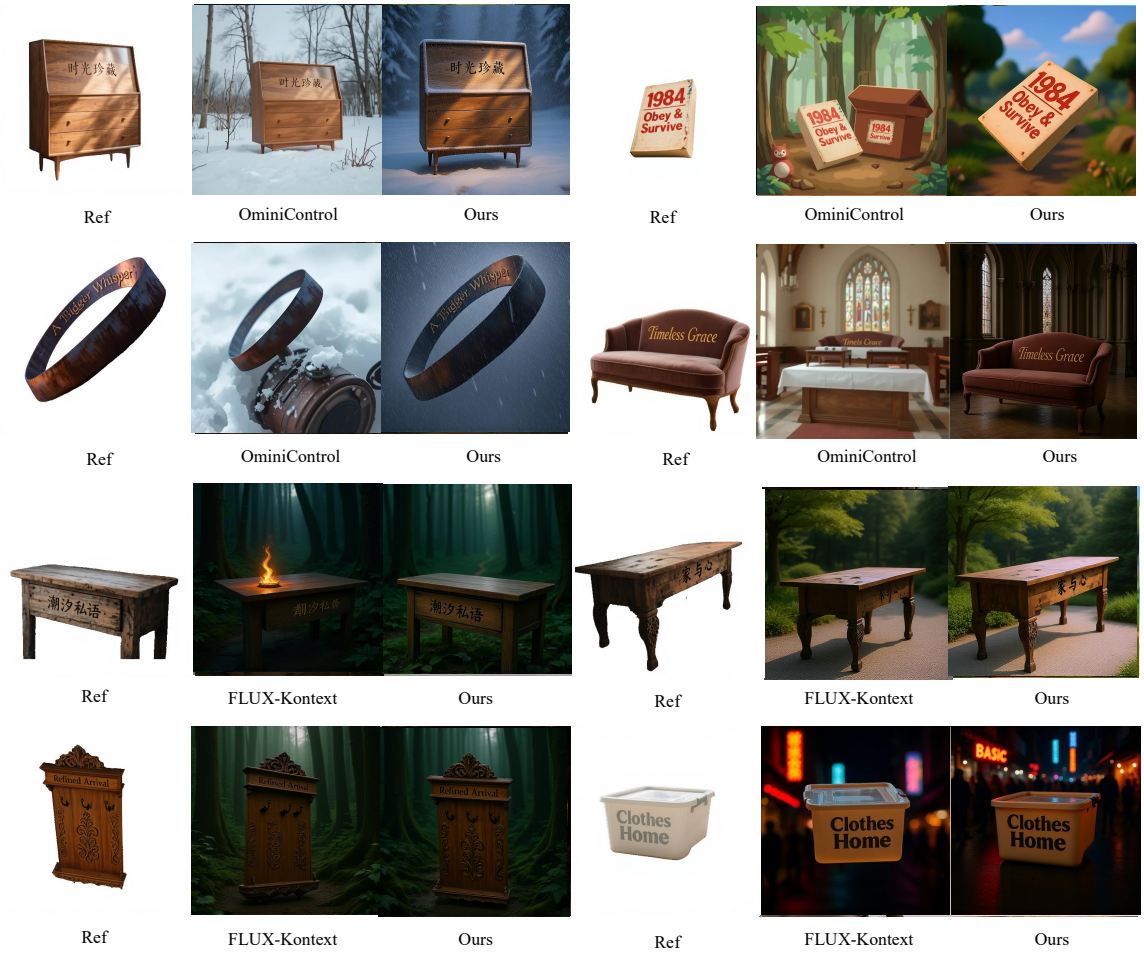

Figure 4. **Qualitative comparison with baseline methods.** Each example shows the reference image (Ref) alongside the outputs from baseline models (OminiControl, FLUX-Kontext) and our method. Compared to the baselines, our method demonstrates superior performance in generating images that are more consistent with the text prompt while more faithfully preserving the subject's identity and intricate textual details.

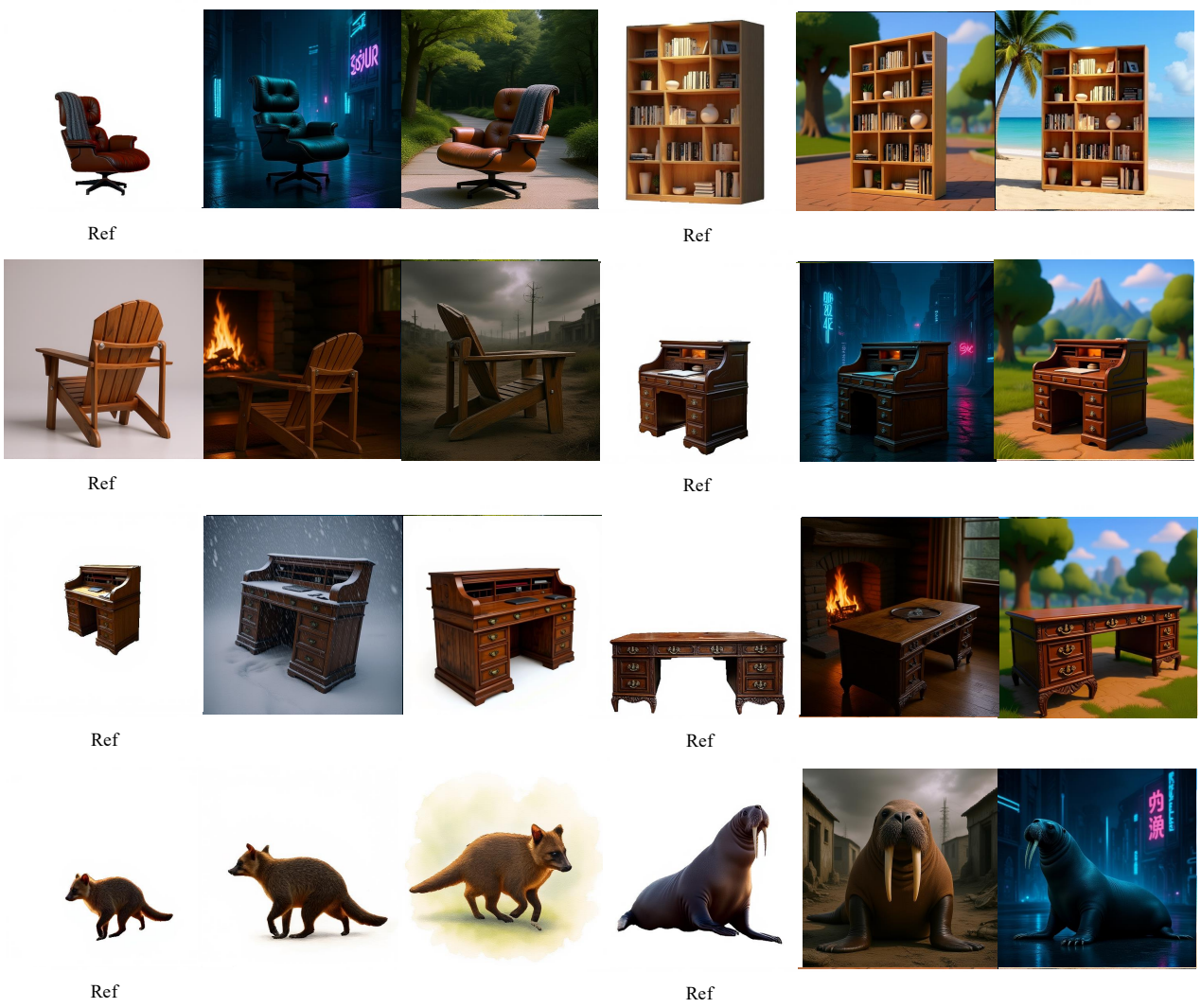

Figure 5. **Additional qualitative results on text-free objects.** Each group displays a reference image (Ref) and two different scenes generated by our model. These examples showcase the model's ability to robustly place the source object into diverse new environments and styles while maintaining its core identity features.

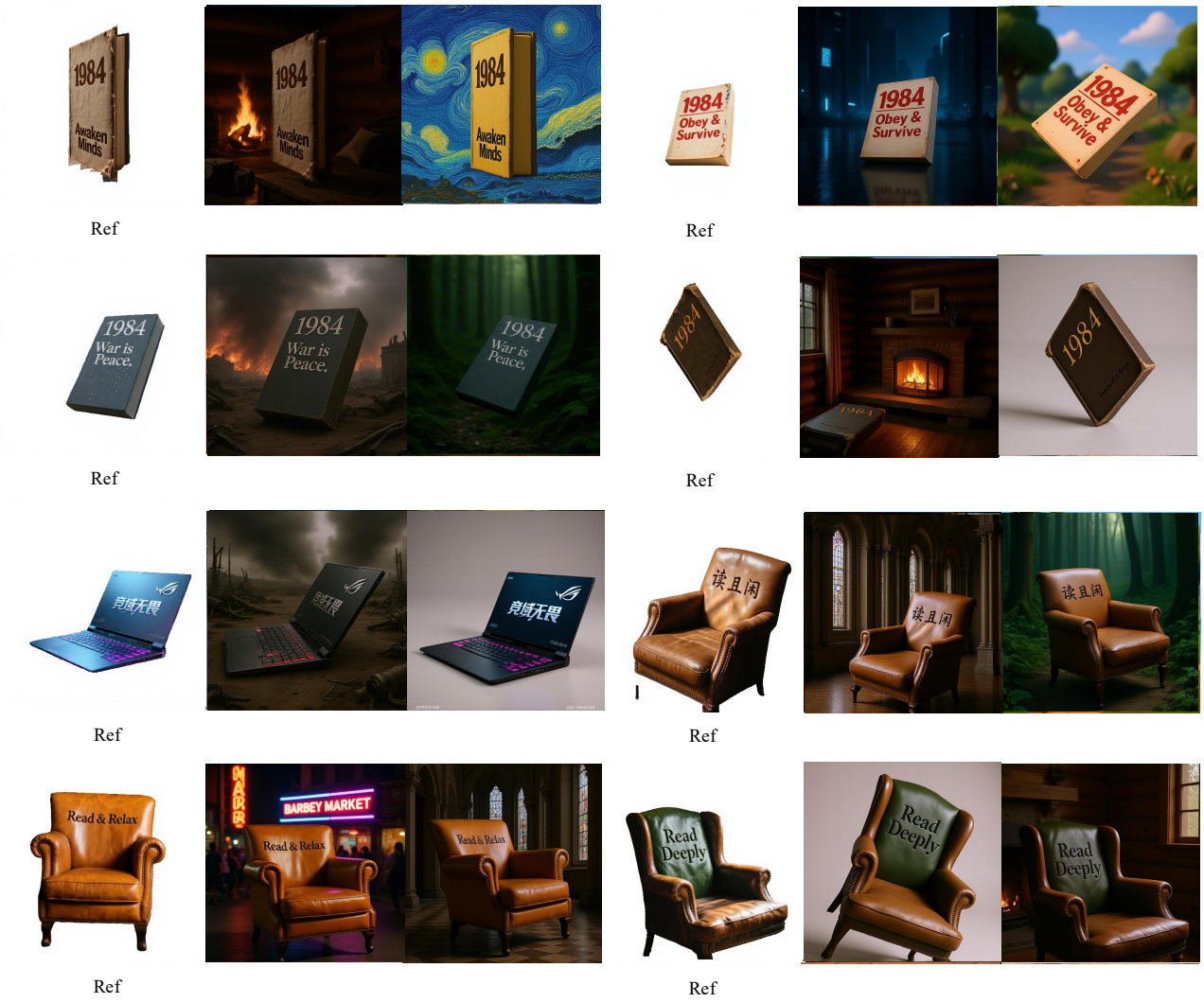

Figure 6. **Additional qualitative results on objects with text.** Each group shows a reference image (Ref) with text and two outputs generated in new scenes. Our method successfully preserves the legibility, style, and content of the text on the object even under significant changes in background and lighting, a key contribution of our work.

Figure 7. **Additional qualitative results on objects with text.** Each group shows a reference image (Ref) with text and two outputs generated in new scenes. Our method successfully preserves the legibility, style, and content of the text on the object even under significant changes in background and lighting, a key contribution of our work.

B. Dataset Examples

This section provides examples from the datasets used to train our model. We augmented the general-purpose ‘Subject200k‘ dataset and constructed the specialized ‘TextingSubject100k‘ dataset to handle text rendering tasks.

Figure 8. **Examples from the augmented ‘Subject200k‘ dataset.** These image triplets are created by taking a source image and applying Bagel to generate novel, rotated views. This data augmentation strategy helps the model learn a robust representation of object identity across different viewpoints.

Figure 9. **Examples from our custom ‘TextingSubject100k‘ dataset.** Each triplet shows an object with text, with its views rotated by Bagel. We employ a strict OCR filtering process to ensure the text remains consistent and legible across views, providing high-quality training data for our text-aware model.