CLiC: Concept Learning in Context

Supplementary Material

A. Additional Results

In this section, we present supplementary results to further demonstrate the capabilities and effectiveness of our method. We provide additional qualitative examples of concept transfer and generation, as well as extended comparisons with baseline methods. Moreover, we report additional quantitative results from user studies, investigate the impact of text prompts and input mask choice, showcase multi-concept learning, and offer an additional showcase of the cross-attention guidance mechanism. These additional results highlight the robustness and versatility of our approach in various scenarios.

A.1. Qualitative Results

Fig. 6 showcases additional results of our concept transfer and generation applications. The settings employed for concept transfer and generation are consistent with those outlined in Sections 3.2 and 4.1. Evidently, our method successfully learns concepts from a variety of objects and utilizes these concepts for image editing and generation.

A.2. Comparison

In Fig. 7, we present additional comparison results alongside the four baselines previously introduced in Section 4.2. It is clearly demonstrated that our in-context concept learning approach exhibits superior proficiency in learning and transferring concepts.

A.3. Quantitative Results

We conducted a user study for our ablated results demonstrated in Table 1. Furthermore, we ran an additional user study for the generation task, demonstrated in table 2. For both of these additional user studies, we followed the procedure described in the paper. For the generation task, the scores are out of 3, and for the ablation, the scores are out of 4 (higher better).

Table 1. **Ablations user study**.

no ℓ_{att}	no ℓ_{con}	no ℓ_{RoI}	Ours
2.49	1.67	2.04	**3.79**

Table 2. **Generation user study**.

CustDiff	BAS	Ours
1.43	1.83	**2.76**

A.4. The Efficacy of Text Prompts

As demonstrated in Fig. 1, varying the prompt in the generation phase effectively alters the generation results (a wooden chair vs iron chair with v* style).

Figure 1. The efficacy of text prompts.

A.5. Comparison with ReVersion

ReVersion [2], which focuses on learning the relation between two objects, can be applied to learn a pattern. We performed a visual comparison illustrated below. While, it does a reasonable job, it lacks spatial and in-context awareness.

Figure 2. Comparison with ReVersion.

A.6. Multi-Concept Learning and Transfer

We show an additional example of learning multiple concepts from a single object, depicted in Fig. 3.

Figure 3. Multi-Concept Learning and Transfer.

A.7. Choice of Input Mask

Our algorithm is robust to the location and size of the mask. To show this capability, we have scaled the mask in Fig. 4 below and learned the same concept. The results show

that the method is still capable of learning the concept and transferring it successfully.

Figure 4. Choice of input mask.

A.8. Cross-Attention Guidance

Fig. 5 shows additional results depicting the effect of cross-attention guidance, where increasing the guidance step size strengthens the presence of the concept.

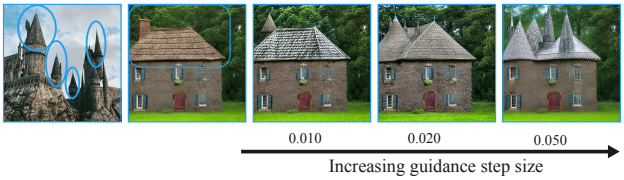

Figure 5. Cross-attention guidance.

B. Additional Training Details

B.1. Data Augmentation Strategies

To enhance the robustness of our approach, we incorporated several data augmentation techniques during the training process. These include implementing random grayscaling to reduce dependence on color features, and preventing overfitting to specific colors. We also applied random horizontal flipping to introduce pose diversity, as well as zooming in and out to vary the scale. To address different color intensities and contrasts, we also employed color jittering.

B.2. Standardized Prompt Templates

For consistency and to prevent the impact of prompt manipulation, we defined a fixed prompt template and used that for all our experiments. Throughout the Concept Learning phase, we utilized a standardized prompt template: "A OBJECT with [v*] style". This uniformity enables effective concept learning and encoding within the [v*] token.

During zoom-in/out data augmentation, the prompt format was dynamically adjusted to reflect these changes. For instance, a zoom-out augmentation led to a prompt alteration to "A OBJECT with [v*] style, zoomed-out".

To maintain equitable comparisons, these augmentations and prompt adjustments were consistently applied across all baseline methods.

B.3. Scheduler Selection

We opted for the DDIM [4] scheduler for both concept learning and transfer phases, due to its efficiency, speed, and simplicity. A maximum of 50 timesteps ($T = 50$) was consistently used in all generation and editing tasks.

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 4

[2] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 1

[3] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 4

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 2

[5] Luming Tang, Nataniel Ruiz, Chu Qinghao, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, and Michael Rubinstein. Realfill: Reference-driven generation for authentic image completion. *arXiv preprint arXiv:2309.16668*, 2023. 4

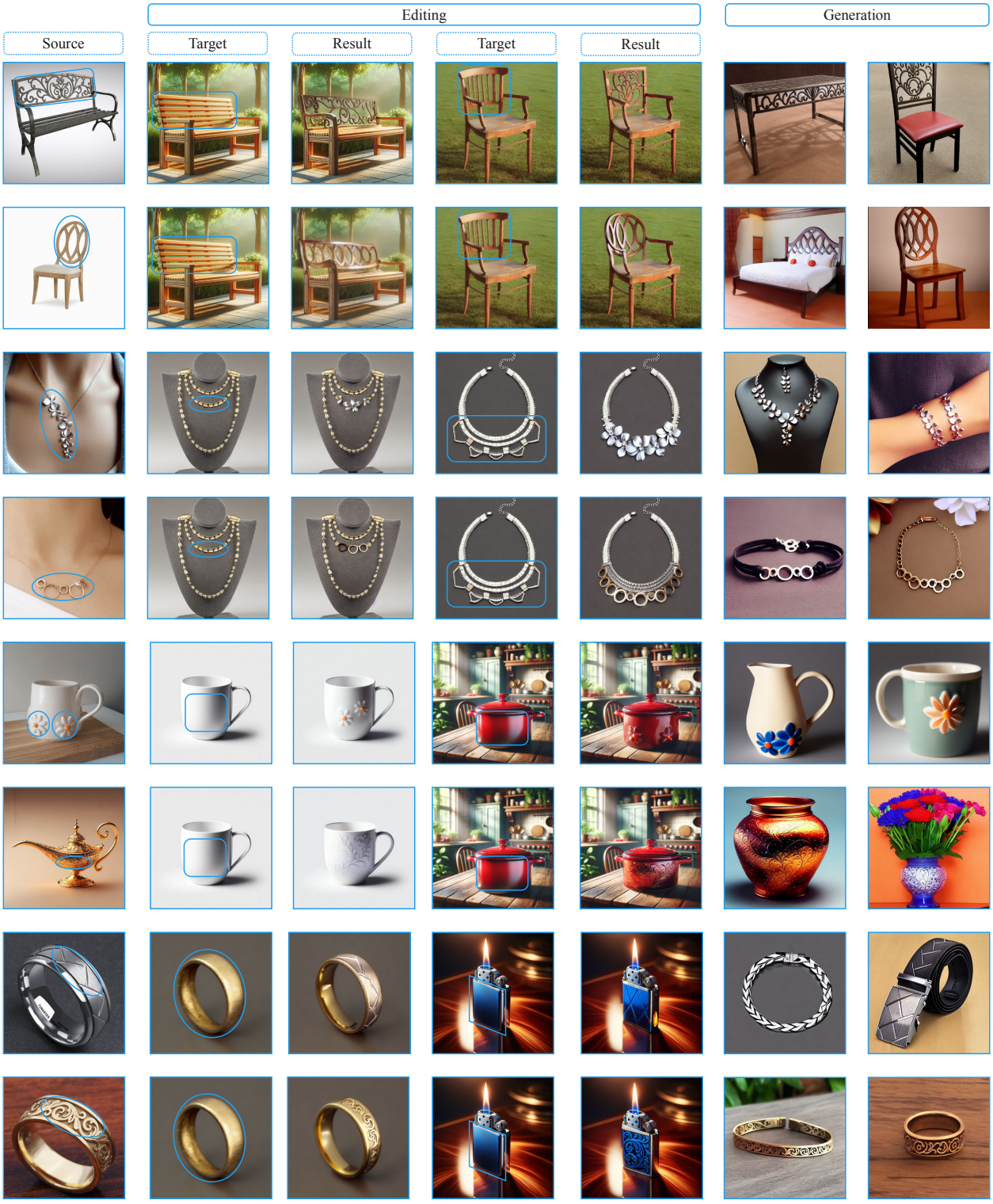

Figure 6. **Additional editing and generation results**. We have transferred the concept from the source to two targets in each row. We also used the same concept for generation (the last two images in each row).

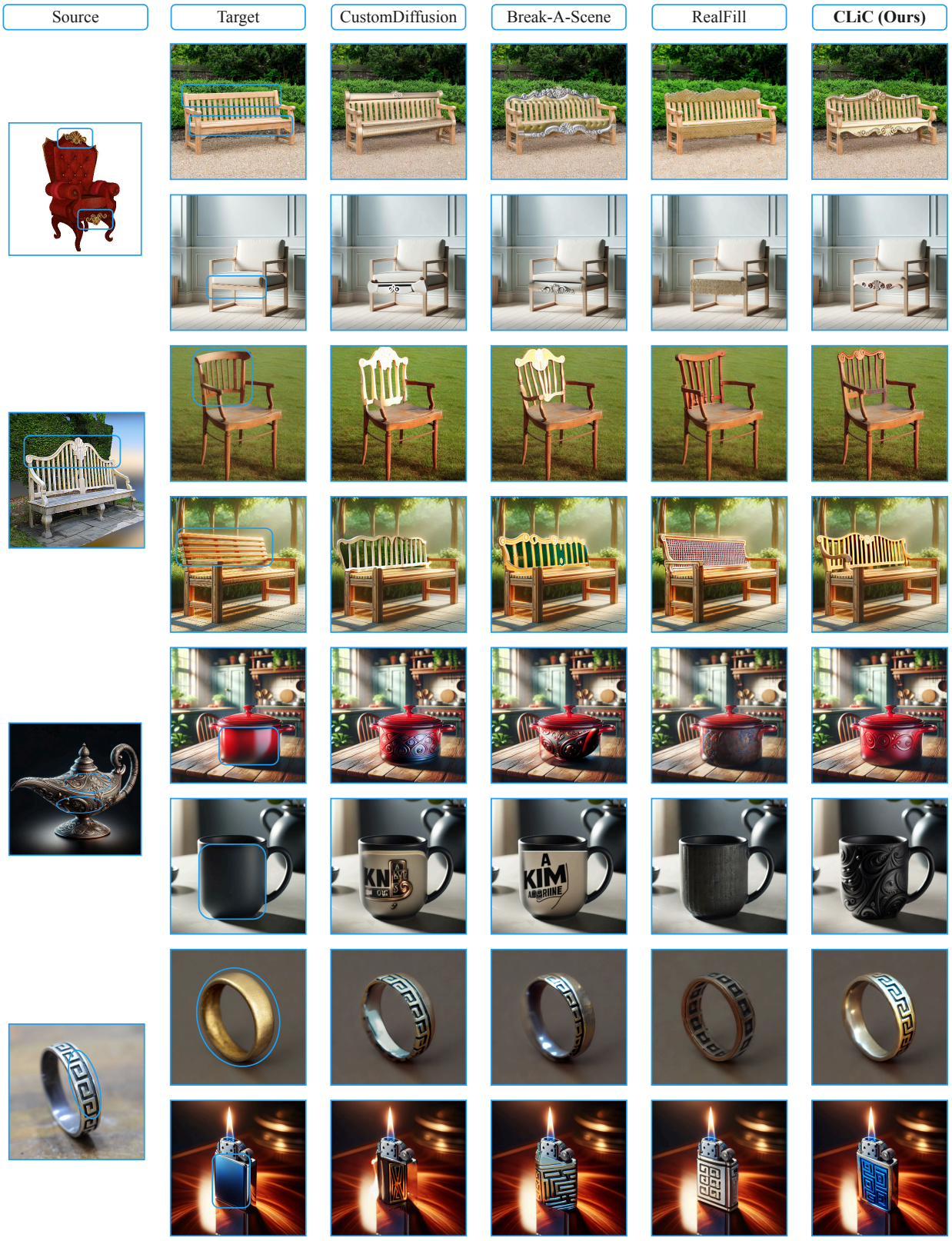

Figure 7. **Additional comparisons**. We further compare our concept transfer method with CustomDiffusion [3], Break-A-Scene [1], and RealFill [5].