CLiC: Concept Learning in Context

Mehdi Safaee[1] Aryan Mikaeili[1] Or Patashnik[2] Daniel Cohen-Or[2] Ali Mahdavi-Amiri[1]

[1]Simon Fraser University [2]Tel Aviv University

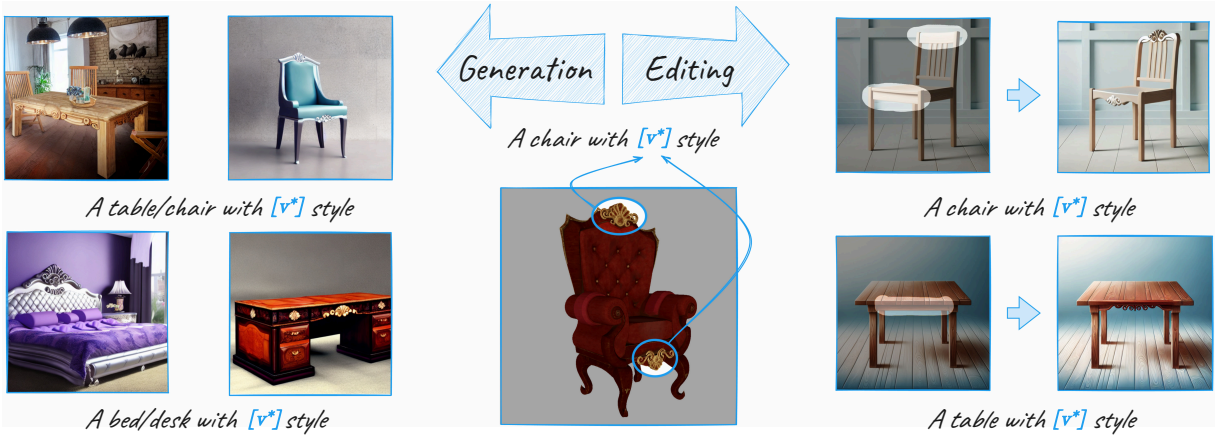

Figure 1. Given an object in an image (e.g., the red chair in the middle), we learn an in-context token for a concept embedded in this object (e.g., an ornament of a chair). This token can then be used to generate images with new objects embodying the same concept (left) or to transfer the concept to given target objects, while maintaining their structure (right). Project page: https://mehdi0xc.github.io/clic

Abstract

This paper addresses the challenge of learning a local visual pattern of an object from one image, and generating images depicting objects with that pattern. Learning a localized concept and placing it on an object in a target image is a nontrivial task, as the objects may have different orientations and shapes. Our approach builds upon recent advancements in visual concept learning. It involves acquiring a visual concept (e.g., an ornament) from a source image and subsequently applying it to an object (e.g., a chair) in a target image. Our key idea is to perform in-context concept learning, acquiring the local visual concept within the broader context of the objects they belong to. To localize the concept learning, we employ soft masks that contain both the concept within the mask and the surrounding image area. We demonstrate our approach through object generation within an image, showcasing plausible embedding of in-context learned concepts. We also introduce methods for directing acquired concepts to specific locations within target images, employing cross-attention mechanisms, and establishing correspondences between source and target objects. The effectiveness of our method is demonstrated through quantitative and qualitative experiments, along with comparisons against baseline techniques.

1. Introduction

Consider the problem of transferring an ornament from one image of a chair onto another image of a different chair, even if the chairs are in different orientations (see Fig. 1). It is evident that a straightforward image-space cut-and-paste operation is insufficient here. Moreover, attempting to model the ornament from a single perspective and accurately pasting it onto the other image is a complex task, one that currently stands as a daunting challenge.

An alternative approach involves harnessing the recently developed domain of visual concept learning [2, 10, 11, 22, 30] that allows learning the visual concept from a source image and subsequently applying it to a target image. While it does not provide an exact, one-to-one transfer of the ornament, it does offer a way to transfer the overall concept. Yet, plausibly learning a local concept from a single image and applying it in a specific location of an object in the target image is challenging due to the lack of context (see Fig. 1).

Avrahami et al. [4] have recently presented a technique called "Break-A-Scene" wherein they learn local concepts from a single image and then apply them within a generated image through text-to-image machinery. This technique can be seemingly applied to our local concept learning problem as well. However, as we shall show in the fol-

lowing, our specific task necessitates learning local visual concepts within the shape's context rather than in isolation. Our learned local concepts are intrinsically tied to the objects in which they are embedded. The method we present in this paper addresses the intricate challenge of in-context concept learning, specifically tailored to our requirements.

To learn a visual concept in-context, we apply a personalization method that learns a token v^*, where a mask defines the spatial region of the acquired concept (e.g., ornament/window). Rather than applying the losses only under the given mask, we compute a loss with a non-binary mask, that is, a soft mask that considers both the in-mask concept and the out-mask portion of the image. Fig. 2 shows two examples generated with a simple text prompt, "A chair/house with v^* style" where one is with the in-context learned concept (the ornament/windows in the figure), and one without. As can be seen, the in-context visual concept learning successfully embeds the concept only in the expected region of the generated results.

We show that the acquired concept can be directed to a specific location within a given target image through the optimization of cross-attention layers and the establishment of correspondences between source and target objects. We further present an automated process for identifying common concepts when multiple objects embodying a particular concept are available, which removes the necessity to manually choose the concept in the source image.

We demonstrate the efficacy of the method via numerous results and multiple quantitative and qualitative experiments and comparisons against baseline methods. We also show the necessity of having each component of the method through a series of thorough ablation studies.

2. Related Work

The field of text-conditioned image generation [5, 28, 29, 32] has recently advanced significantly by combining the power of diffusion models [15, 35–37] and large-scale text-image datasets [33]. These advancements have had great contributions to the area of content creation, showcasing the capacity of these models to produce captivating visual content, enabling a multitude of creative visual tasks through image generation and editing [6, 14, 19, 24, 25, 41]. One such task is to utilize user-defined concepts [10, 30] to accommodate *personalization* [2, 13, 18, 22, 27, 43, 44], empowering users to craft expressive content that seamlessly blend subjects and artistic styles, often requiring just a small collection of concept-exemplifying images.

The first attempts to address personalization were Textual Inversion [10] and DreamBooth [30]. In both works, given multiple images of a single concept, a text token dedicated to that concept is learned. However, while the former freezes the weights of the diffusion model UNet, the latter optimizes the UNet, showing better reconstruction and generalization capabilities at the expense of additional time and memory consumption. Custom Diffusion [22] approaches this problem by optimizing only the cross-attention layers of the diffusion UNet, while OFT [27], LoRA [17], and SVDiff [13] restrict the parameter updating for more efficient and well-behaved optimization. Similarly, PerFusion [40] introduces a key-locking mechanism along with rank 1 updating for faster, better, and less memory-consuming personalization. More recently, several works have improved personalization by decreasing runtime and focusing on a single input image [9, 11, 18, 31, 34]. Other works learn multiple concepts [4, 13, 22] or extend the text embedding space of the diffusion model [2, 43].

Figure 2. By learning a concept in-context, the comprehension of the concept extends beyond its visual attributes to encompass its relationship with the surrounding context. In this example, when learning the ornaments and windows in-context, they are placed in similar semantic locations in the generated images as in the source image. Conversely, when ornaments are learned without the context, they may be dispersed randomly across the chair and house.

In Break-A-Scene [4], multiple concepts are learned given a single image and user-defined masks. Specifically, the concepts are learned by using a masked diffusion loss and restricting the cross-attention maps of the learned tokens to the input masks. Unlike their work, our method addresses the in-context concept learning, specifically tailored to our requirements. Also, concurrent to our work, RealFill [39] tackles the problem of personalized image inpainting and outpainting by fine-tuning an inpainting diffusion model [29] on a collection of input images. However, when RealFill is employed for our concept transfer task, the relative size of the concept to the base object is not maintained, and geometric details are compromised.

Recently, many works have utilized the intermediate features of diffusion models for image editing [7, 12, 14, 25, 41], controlled image generation [8, 16, 26], and image understanding [1, 20, 23, 38, 45]. Prompt2Prompt [14] shows that by manipulating the cross-attention layers of the diffu-

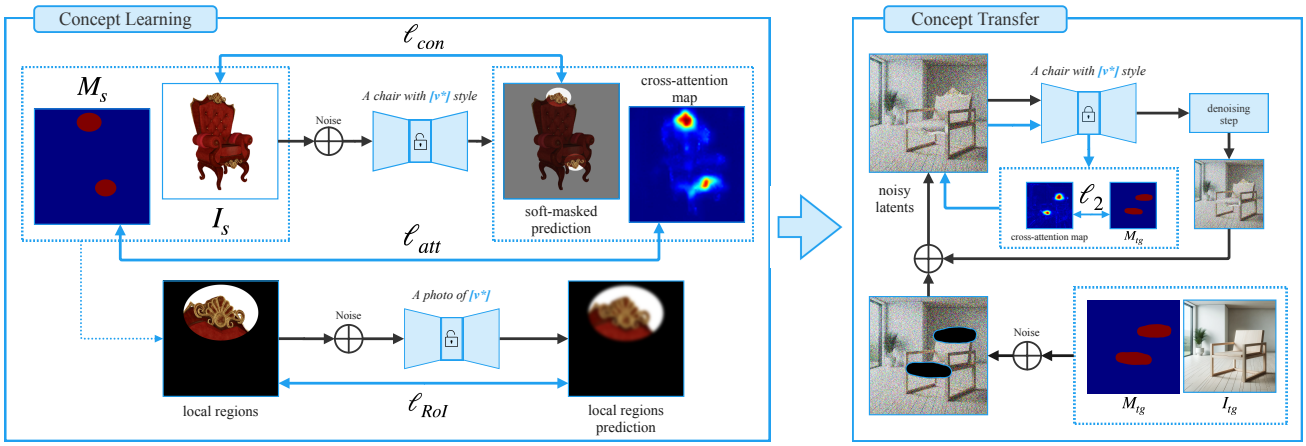

Figure 3. **In-Context Concept Learning**: given image I_s and a binary mask M_s, we learn v^* for a concept outlined by mask M_s in the context of a base object. Here, the concept is the ornament, and the base object is a chair. Three loss functions are utilized to optimize v^* and fine-tune the diffusion model. $\boldsymbol{\ell_{con}}$ uses a soft-masked diffusion loss to learn the pattern in context. $\boldsymbol{\ell_{att}}$ ensures that the token focuses only on the pattern region by restricting the attention maps of v^* to M_s. By employing a text prompt that is specified for v^*, $\boldsymbol{\ell_{ROI}}$ enhances the reconstruction of the concept by focusing on a local region through masking I_s. **Concept Transfer**: given image I_{tg}, mask M_{tg} defining the area of edit, and a user-defined text-prompt containing v^* optimized in the concept learning step, we add noise to the latent of I_{tg} and denoise it with the fine-tuned diffusion model obtained from the concept learning step. At each denoising step, we blend the output of the diffusion model with the masked input to preserve the out-of-mask regions. We also have a cross-attention guidance to enhance the presence of the pattern in the final output.

sion model, it is possible to edit a certain semantic region of an image. Attend-and-Excite [8] ensures that the diffusion model attends to every subject in the text prompt by manipulating the cross-attention maps of the subjects in the generation process. More recent methods [1, 20] demonstrate that by optimizing text tokens and cross-attention maps, it is possible to establish semantic correspondence or segmentation. In our work, we also use the cross-attention maps of the diffusion model to localize the learned tokens to the acquired visual concepts in the source image and automatically place them correctly in the target image.

3. Method

Given a source image, denoted as I_s, a user-specified prompt P_s, and a learned or user-provided mask M_s that marks the Region of Interest (RoI) within the source image, our objective is to learn the concept (e.g., an ornament) in the RoI. To achieve this objective, we build on previous works that employed text-to-image diffusion models for the task of personalization. Such works either optimize a text token v^*, fine-tune the pretrained text-to-image model, or their combination. In our work, we opt to use Custom Diffusion [22], optimizing a text token v^* and simultaneously fine-tuning the cross-attention layers of the text-to-image model. To learn the concept in-context (Section 3.1), we employ multiple loss functions, encouraging the diffusion model to reconstruct the learned concept in analogous contexts but under varying conditions and poses. After learning the concept, we can either generate images that contain it, or edit a target image I_{tg} to portray it in a given RoI. The RoI in the target image is determined either via our Diffusion-Based RoI Matching Algorithm proposed in Section 3.3 or directly provided by the user.

3.1. In-Context Concept Learning

To acquire in-context concepts from the source image, we optimize token v^* and simultaneously fine-tune the cross-attention layers of a pretrained T2I diffusion model as done in Custom Diffusion [22]. We employ three loss functions to ensure effective concept learning and precise in-context generation. ℓ_{att} helps the model to focus on the RoI. $\ell_{context}$ facilitates in-context concept learning. Although the concept is learned in-context, ensuring that the token possesses knowledge of the concept's inclusion within a larger object, we employ ℓ_{RoI} to safeguard against overfitting the concept to a particular object in the source image. This approach enhances the concept's ability to generalize and transfer to unseen objects, even those from different categories. Additionally, ℓ_{RoI} aids in acquiring a more nuanced understanding of the concept's geometric attributes.

Given the source image I_s, its corresponding binary mask M_s, and a text prompt P_s, we first encode the image and prompt to obtain latent image x_0 and text embedding c. Thereafter, by randomly sampling a timestep t from the interval $[1, T]$ and a noise ϵ, we construct a noisy latent x_t. We then employ the diffusion model to get $\epsilon_\theta(x_t, c, t)$

while also extracting cross-attention maps for the token v^* from the decoder layers of the UNet structure. The cross-attention loss is then computed as:

$$\ell_{att} = \mathbb{E}_{(x_t,t)} \left[\| CA_\theta \left(v^*, x_t \right) - Resize(M_s) \|_2^2 \right], \quad (1)$$

where $CA_\theta \left(v^*, x_t \right)$ denotes cross-attention maps between token v^* and x_t averaged over the cross-attention layers of the upsampling blocks and $Resize(M_s)$ is the resized version of M_s that matches the shape of cross-attention maps.

For the context loss, we aim to make the model focus on in-context reconstruction of the concept in the RoI while simultaneously forcing the model to focus on the proper scale and placement of this concept. To achieve this, we employ a soft-weighted version of M_s:

$$M_{soft} = \alpha + (1 - \alpha) M_s, \quad (2)$$

where $\alpha = 0.5$. The context loss is then computed as:

$$\ell_{con} = \mathbb{E}_{(x_t,c,t)} \left[\| M_{soft} \odot \left(\epsilon_\theta \left(x_t, c, t \right) - \epsilon \right) \|_2^2 \right]. \quad (3)$$

For RoI loss, we use a more concept-oriented prompt "A photo of v^*" encoded into c^*:

$$\ell_{RoI} = \mathbb{E}_{(x_t,t)} \left[\| \epsilon_\theta \left(M_s \odot x_t, c^*, t \right) - \epsilon \|_2^2 \right] \quad (4)$$

and we finally add all the losses and perform optimization:

$$\ell_{tot} = \ell_{con} + \lambda_{att} \ell_{att} + \lambda_{RoI} \ell_{RoI}, \quad (5)$$

where λ_{att} and λ_{RoI} are empirically set to 0.5.

3.2. Concept Transfer

To transfer the learned concept to new objects while preserving the region outside the Region of Interest (RoI), we utilize masked blended diffusion editing [3]. This involves adding a specific amount of Gaussian noise to the target image to reach the timestep t_{start} of the denoising process. We then begin denoising the image, simultaneously blending the out-of-mask region of the target image at each denoising step. Additionally, adapted from Attend-and-Excite [8], we introduce cross-attention guidance to improve control over the strength of the edit. In this process, we gradually optimize the latents so that the cross-attention map of the v^* token increases in the RoI and decreases elsewhere.

Blended Diffusion Editing. Given a target image I_{tg} and its corresponding mask M_{tg}, we aim to modify segments within $M_{tg} \odot I_{tg}$. First, the target image is encoded, and x_{tg} is obtained, then an initial timestep t_{start} is chosen ($5 \leq t_{start} \leq 15$). Next, we add $T - t_{start}$ levels of noise to x_{tg} to obtain $x'_{t_{start}}$, then, at each timestep $0 < t \leq t_{start}$, blended output x'_t is computed as:

$$x'_t = M_{tg} \odot x_t + (1 - M_{tg}) \odot x'_{start}. \quad (6)$$

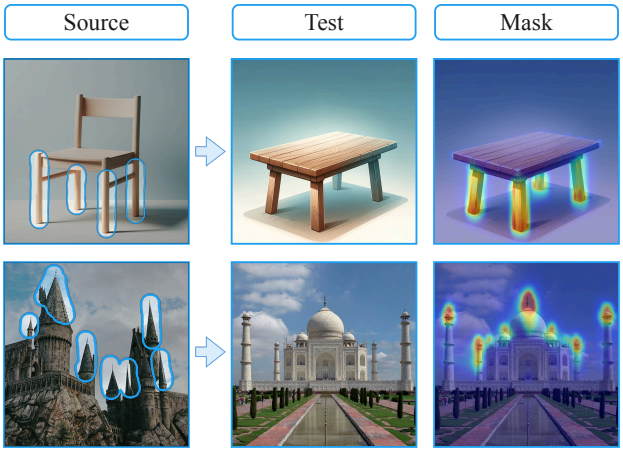

Figure 4. Illustration of the automated masking process on target images employing the proposed RoI-Matching technique, leveraging a predefined source mask.

Cross-Attention Guidance. After obtaining x'_t, we extract the attention maps of the v^* token $CA_\theta(v^*, x'_t)$. Then, we update x'_t according to Equation 7 to enhance the strength of the attention maps of v^* within M_{tg}:

$$x''_t = x'_t - \eta \nabla \mathbb{E}_l [\| CA_\theta \left(v^*, x'_t \right) - M_{tg} \|_2^2]. \quad (7)$$

Here η is the step size of the guidance. changing this parameter controls the strength of the edit in the RoI. Finally, we denoise x''_t through the UNet.

3.3. RoI Matching

Automatic Target Mask Extraction. Mask extraction on the target image according to the source input mask can be automated. The idea is to learn a new token w^* to the text encoder, initialized with the already optimized v^* and optimizing it by minimizing our attention loss, ℓ_{att}, using the prompt "a w^* region of an OBJECT", with OBJECT being the base object in the source image. After 500 steps of optimization, we apply the new token on the target or other source images and execute the denoising process, extracting the attention maps of the token w^* as the target masks. By doing this, the model acts as a segmentation method that segments the corresponding part of the target or source images. In Fig. 4, we demonstrate that this automatic masking technique works well both for in-domain and cross-domain scenarios. Notably, this process is quite fast since we only fine-tune the newly added token.

Automatic Source Mask Extraction. When multiple source images sharing the desired concept exist, it can be automatically identified. To do so, we add a token w^* to the text encoder and optimize its embedding and the cross-attention modules of the diffusion UNet by minimizing the diffusion loss, given the prompt "An OBJECT with w^*

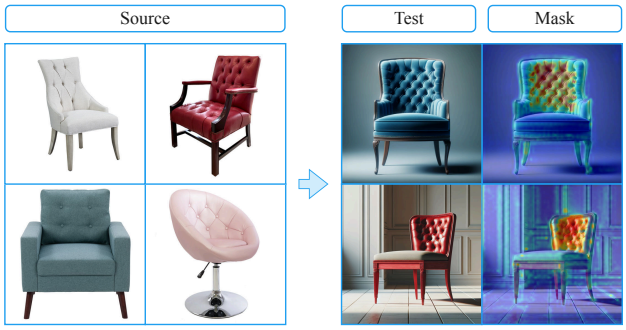

Figure 5. Given several images with a common pattern, our method is able to learn the common pattern and locate it even on a different image with the same pattern.

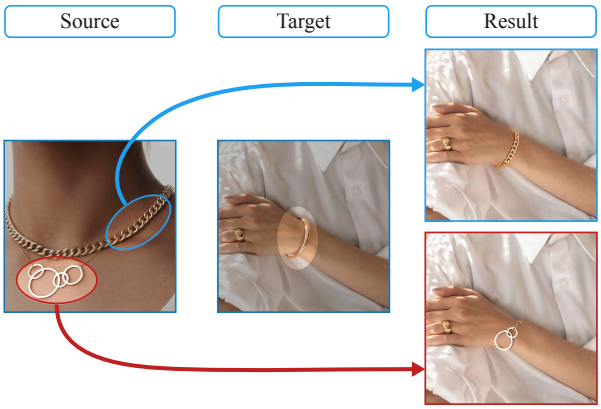

Figure 6. An example of selective pattern extraction: We show that our model can learn distinct patterns from a single image, ensuring that each token captures only its corresponding pattern. Left, we choose two different patterns from a single necklace and transfer them to a bracelet (Right).

style", OBJECT being the class of base object containing the pattern. After 500 steps, we extract the attention maps of w^* and use them as the source mask, and run our concept learning pipeline. This method is effective when multiple images of an object containing the concept exist (Fig. 5) but for unique concepts, it could be simpler to just provide the source mask.

4. Results and Comparisons

Here, we first qualitatively demonstrate the effectiveness of our method in learning concepts in-context. We show that our learned concepts can be used for generation and transferred across images. We compare our method with multiple customization methods such as Custom Diffusion [22], Break-A-Scene [4], and RealFill [39], showing the superiority of our method (Section 4.2). We also provide a user study, confirming the effectiveness of our method compared to these baselines. Finally, we ablate the components of our pipeline to justify our design (Section 4.3).

In all of our experiments, we use StableDiffusion v1.4 from the diffusers library [42]. We run our in-context concept learning for 500 steps, taking approximately 3 minutes on a single Nvidia RTX3090 GPU. We use the Adam [21] optimizer with learning rate $1e^{-5}$.

4.1. Qualitative Results

Editing. Our method can successfully learn various concepts and transfer them to objects of the same or different class in an image. Fig. 6 shows that our method can learn individual concepts from a single image without color and shape information from other concepts leaking into the token. Fig. 11 illustrates examples of various classes. Note that the learned concepts are blended nicely in the target image, attaining the target's texture and color style even when the target domain is very different from the source domain, such as the cartoonish car and house examples. This demonstrates that our approach does not suffer from overfitting to the concept or the content of the source image, and it reaffirms that a simple copy-and-paste algorithm is not suitable for achieving our objective.

Generation. To generate an object containing the learned concept, we employ a two-stage generation strategy. Starting from a Gaussian noise, for the first $t_s = 5$ steps of denoising, we use the un-modified UNet with the text prompt "a photo of an OBJECT" where OBJECT is the object we want to generate. After the t_s steps, we substitute the UNet with our fine-tuned model and use the text prompt "a photo of an OBJECT, with v^* style".

This way, we leverage the capabilities of the pre-trained diffusion model in generating general realistic images while integrating specific patterns and concept details into the output through our fine-tuned model, which possesses an enhanced understanding of our desired concept. We present our generation results in Fig. 11. Observe that concepts learned from an object (e.g., chair), can be effectively used to generate other objects embodying the same concepts.

4.2. Comparisons

We compare our method against several personalization baseline methods, including Custom Diffusion [22], Break-A-Scene [4], and RealFill [39]. In Custom Diffusion [22], cross-attention blocks, along with token v^*, are optimized for customization by minimizing the unmasked diffusion loss. We run Custom Diffusion in a consistent manner with our setting, with inputs consisting of the source image and the text prompt "an OBJECT with v^* style," where OBJECT denotes the object category (e.g., chair) embodying the concept. Break-A-Scene [4] learns several concepts from a single image using masks indicating different subjects. Similar to our setting, we optimize the cross-attention blocks and token v^* representing a local mask located on the

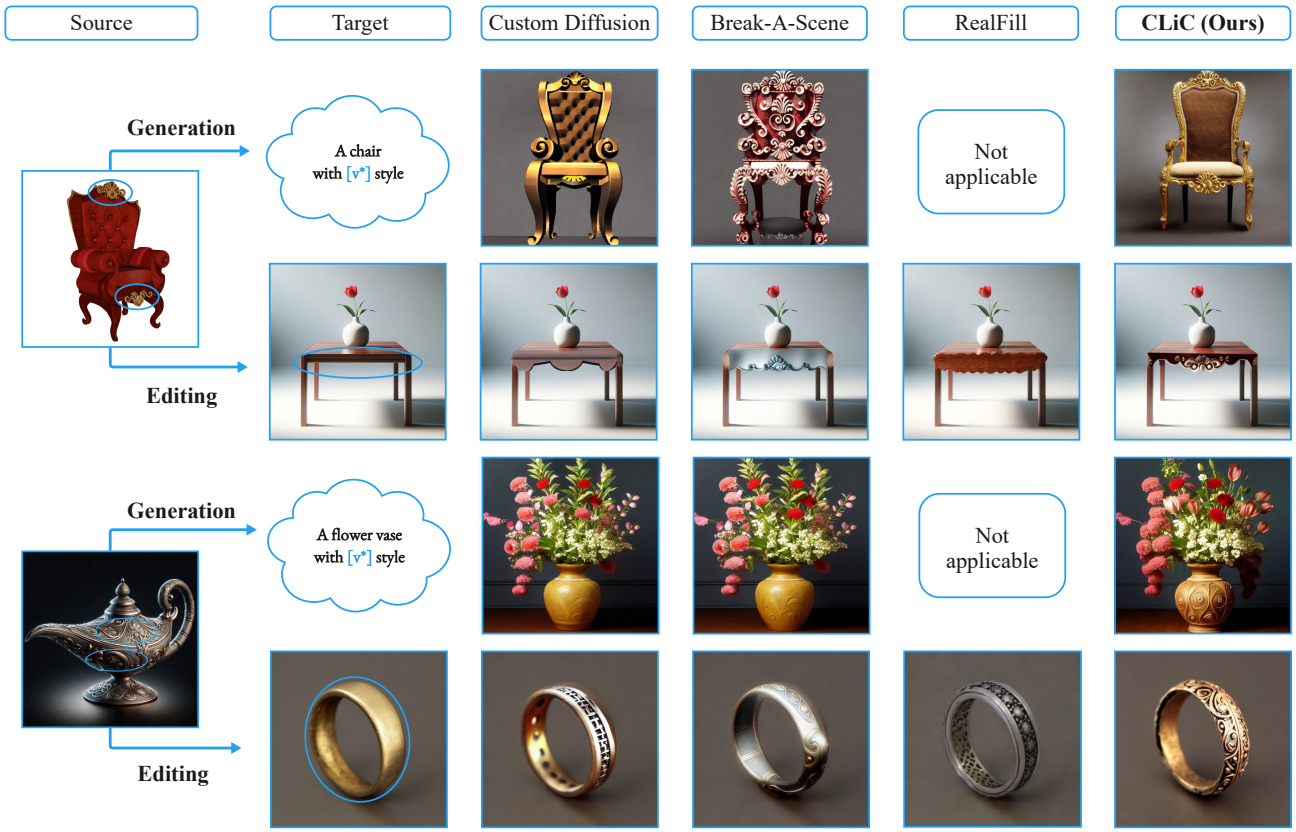

Figure 7. **Comparisons**. Given a source image and concepts of interest (Left), the task is either to generate an object (written in bold (Top)) with that concept or transfer the concept to a target object in another image (Bottom). In comparison with alternative methods, our method clearly remains more faithful to the concept in terms of structure and geometric features in both generation and editing.

concept of interest. We also conduct a comparison against a concurrent work, RealFill [39], designed for personalized inpainting/outpainting. RealFill takes multiple images of a scene as input, randomly applies masks, and refines the Stable Diffusion inpainting model through a process similar to DreamBooth [30]. Our transfer task can be viewed as inpainting. To adapt RealFill to our task, we learn the concept delineated by the mask on the source image and optimize the cross-attention blocks of the Inpainting Stable Diffusion. For transfer, we use the fine-tuned UNet and inpaint the masked regions of the target image using the token acquired from the concept present in the source image.

Qualitative Comparison. In Fig. 7, we present qualitative comparisons with the baselines. Custom Diffusion struggles to capture the concept present in the source images, failing to transfer the concept to the target images effectively. Break-A-Scene exhibits a relatively good understanding of the concept. However, due to the absence of in-context constraints in the concept-learning process, the model learns the pattern as an independent object, failing to transfer the concept as a pattern. This results in unwanted color and geometry artifacts. Similarly, in RealFill, the model learns the token, yet encounters two challenges. First, using the Stable Diffusion inpainting pipeline results in the loss of information masked by the target mask, preventing the model from preserving the geometry and color of the object in the target image (ring in Fig. 7). Second, the absence of in-context learning causes the model to fill the entire mask with the pattern without placing it coherently within the target object (table in Fig. 7).

User Study. We also conducted a user study using 30 pairs of source and target images. Results of our method and three other baselines (depicted in Fig. 7), were presented to 42 participants. The 30 images were divided into two sets, each with a consistent number of object classes (buildings, furniture, jewelry, and kitchenware). Participants ranked the methods based on "edit quality" (accuracy in reflecting the source image concept) and "target preservation" (maintaining the general appearance and color of the target image). Scores were computed by assigning ranks (4 for the top, 1 for the lowest) and averaging over all samples. Our method consistently outperformed the three baselines, as detailed in Table 1. Compared to the second place, RealFill, our method showed a significantly higher score.

Table 1. **User study**. Our method has received a significantly higher score than the alternatives.

Method	Average Ranking (↑)
CustomDiffusion [22]	1.96
Break-A-Scene [4]	2.27
RealFill [39]	2.33
Ours	**3.43**

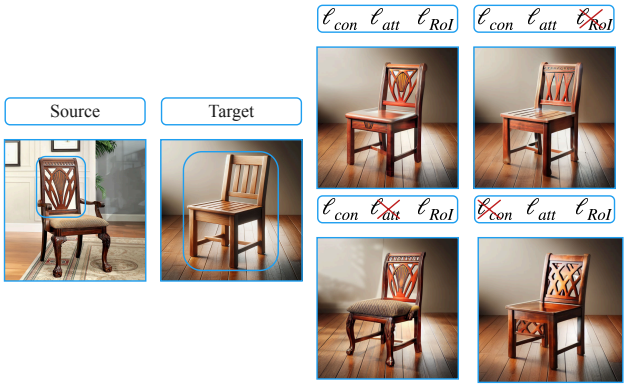

Figure 8. **Ablation on ℓ_{att}, ℓ_{RoI}, ℓ_{con}**. Omitting ℓ_{RoI} causes inaccurate learning of the concept (back seat). Excluding the ℓ_{att} produces unintended or off-target edits (seats and legs). Removing ℓ_{con} leads to the loss of geometric features and structures associated with the concept (back seat) and also results in the transfer of concepts to undesired regions (the transition between two legs).

4.3. Ablation Studies

Loss Ablations. Fig. 8 illustrates how each loss in our approach affects the overall performance when the target region encompasses the entire object. Eliminating ℓ_{RoI} (top right) leads to the loss of geometric and structural patterns from the source concept, particularly noticeable on the backseat. Removal of ℓ_{att} (bottom left) causes undesired alterations on the target, affecting areas such as the legs and seat. The absence of ℓ_{con} (bottom right) leads to loss of geometric details and structures associated with the concept (back seat) and results in the unintended transfer of concepts to undesired regions (transition between two legs); same artifacts illustrated in Fig. 2 for the generation process.

Cross-Attention Guidance. As described in Section 3.2, we use cross-attention guidance to enhance the presence of the concept to the RoI in the target image while also restricting it to the RoI. In Fig. 9, one can observe that by changing the guidance step size η the presence of the concept in the target image can be adjusted.

5. Discussion and Conclusions

We have addressed the challenge of learning and transferring visual concepts between images, focusing on acquiring and applying local visual concepts in-context. Traditional cut-and-paste methods have proven insufficient in these contexts, motivating the exploration of visual concept learning. Our personalization method, which considers both in-mask and out-mask regions of an image, has proven successful in embedding concepts accurately. Precise concept placement has been achieved through the optimization of cross-attention layers and object correspondences, complemented by an automated concept selection process that streamlines the overall workflow.

We have demonstrated the efficacy and versatility of our method, and its capability to learn local concepts for editing and generation. However, we acknowledge certain limitations. Our method may exhibit sub-optimal performance when there is a significant difference in the domain of the target image or the objects for generation compared to the source image (see Fig. 10). Additionally, our optimization process, while effective, is time-consuming and not applicable to real-time applications. We have validated our approach through a diverse set of experiments, quantitative assessment through a user study, a series of qualitative assessments and ablation studies, and comprehensive comparisons with baseline methods. Exploring the potential of our method for 3D concept transfer and geometry editing presents an intriguing avenue for future research.

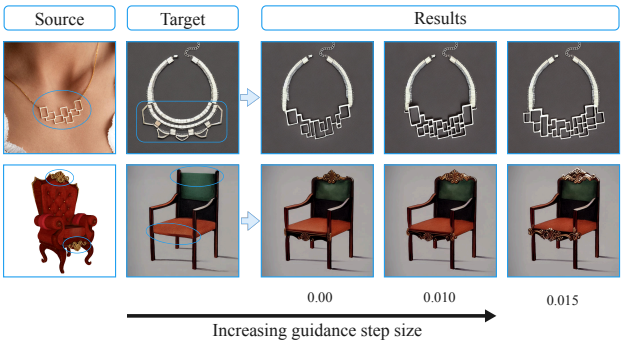

Figure 9. **Cross-Attention Guidance**. By increasing the guidance step size η the presence of the concept is strengthened.

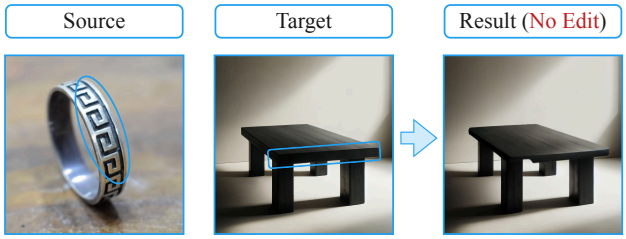

Figure 10. **Failure Case**. When the domain of the source and target images are too different, concept transfer may fail.

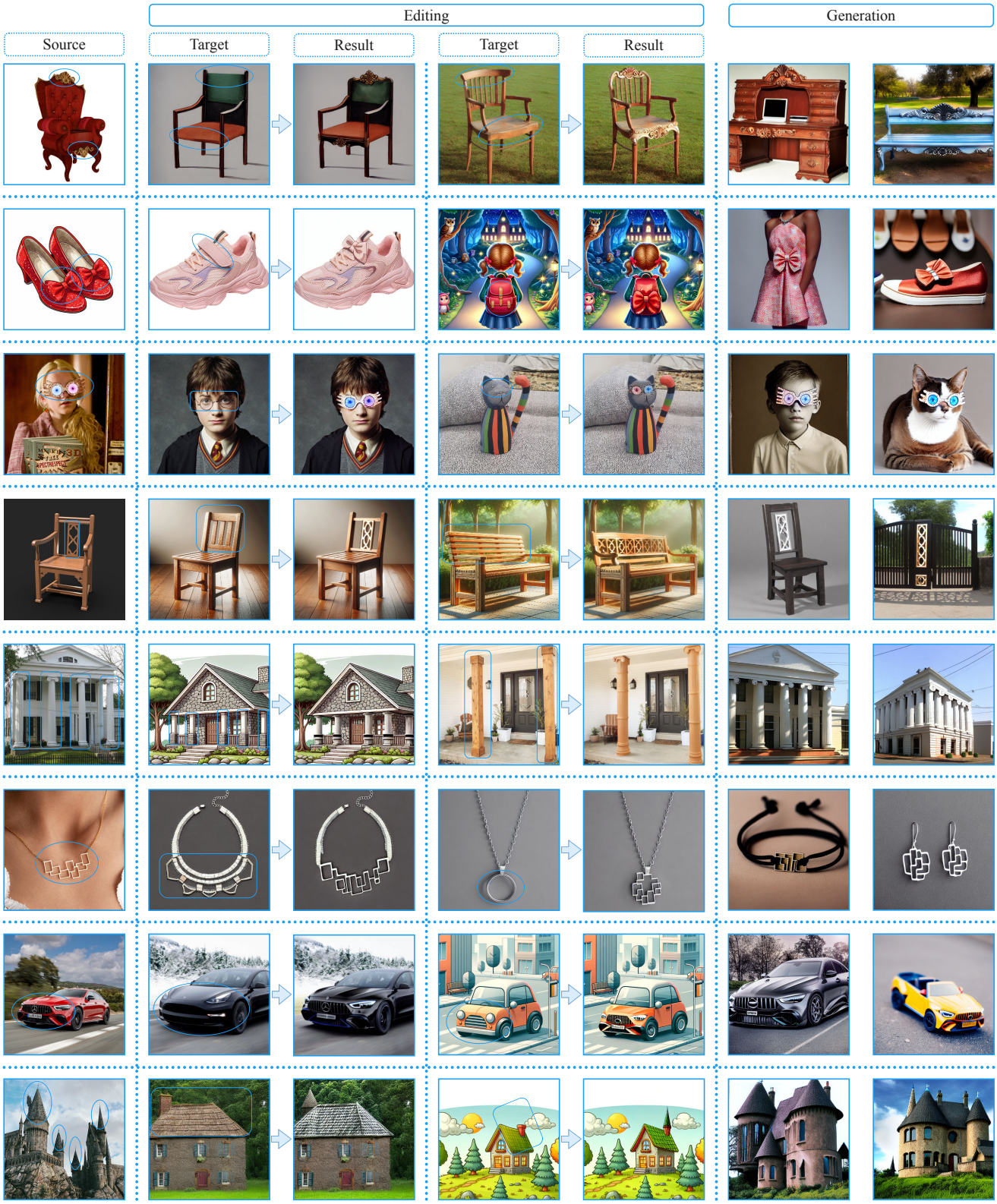

Figure 11. Results of our concept transfer (Middle) and generation (Right). Concepts delineated by blue curves in the source image are learned and transferred to target images at the locations indicated by blue curves (Middle). The same concepts are used to generate various objects in each row (Right). Our method is successful in learning the concept and placing it coherently within the target or generated image.

References

[1] Unsupervised semantic correspondence using stable diffusion. 2023. 2, 3

[2] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization, 2023. 1, 2

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 4

[4] 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. 1, 2, 5, 7, 3

[5] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2

[7] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 2

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 2, 3, 4

[9] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 2

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 1, 2

[11] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models, 2023. 1, 2

[12] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023. 2

[13] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[16] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. *arXiv preprint arXiv:2210.00939*, 2022. 2

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2

[18] Xuhui Jia, Yang Zhao, Kelvin C. K. Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models, 2023. 2

[19] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 2

[20] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me, 2023. 2, 3

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[22] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 1, 2, 3, 5, 7

[23] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. 2

[24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2

[25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. 2023. 2

[26] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[27] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *arXiv preprint arXiv:2306.07280*, 2023. 2

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical textconditional image generation with clip latents, 2022. 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021. 2

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine

tuning text-to-image diffusion models for subject-driven generation. 2022. 1, 2, 6

[31] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models, 2023. 2

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 2

[34] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning, 2023. 2

[35] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 2

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 1

[37] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 2

[38] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2

[39] 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. 2, 5, 6, 7, 3

[40] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[41] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 2

[42] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[43] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation. 2023. 2

[44] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[45] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. 2

CLiC: Concept Learning in Context

Supplementary Material

Here, we provide more results, comparisons, and additional implementation details to better prove the efficacy of our technique.

A. Additional Results

In this section, we present supplementary results of our method. We begin by offering further examples of our concept transfer and generation method, detailed in Section A.1. Subsequently, we include additional comparison results against baseline methods in Section A.2.

A.1. Qualitative results

Fig. 12 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 Figure 13, 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.

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 [36] 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.

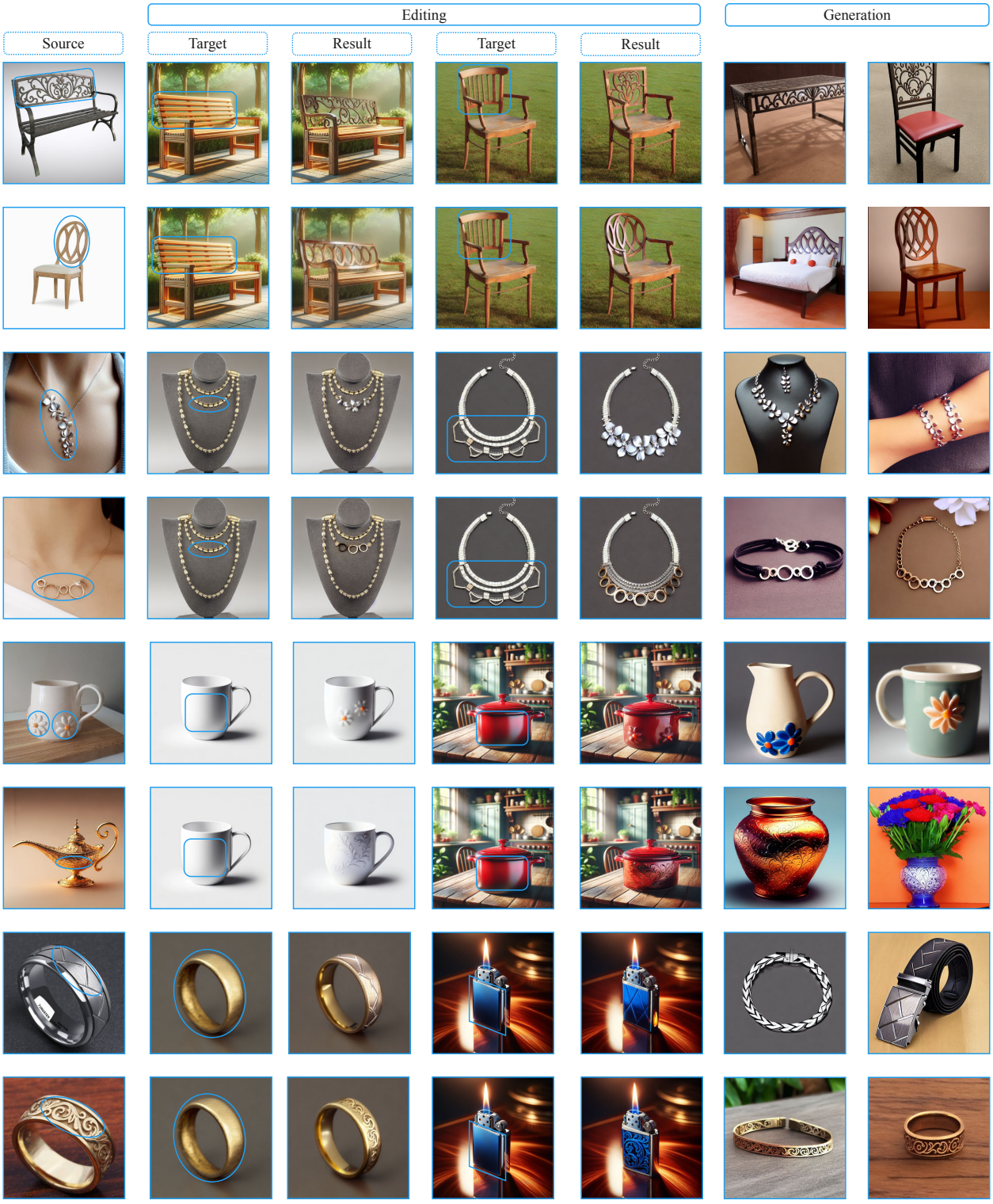

Figure 12. **Additional editing and generation results**. We have transferred the concept from the source to two targets in each row. We also used the same concept for generation (the last two images in each row).

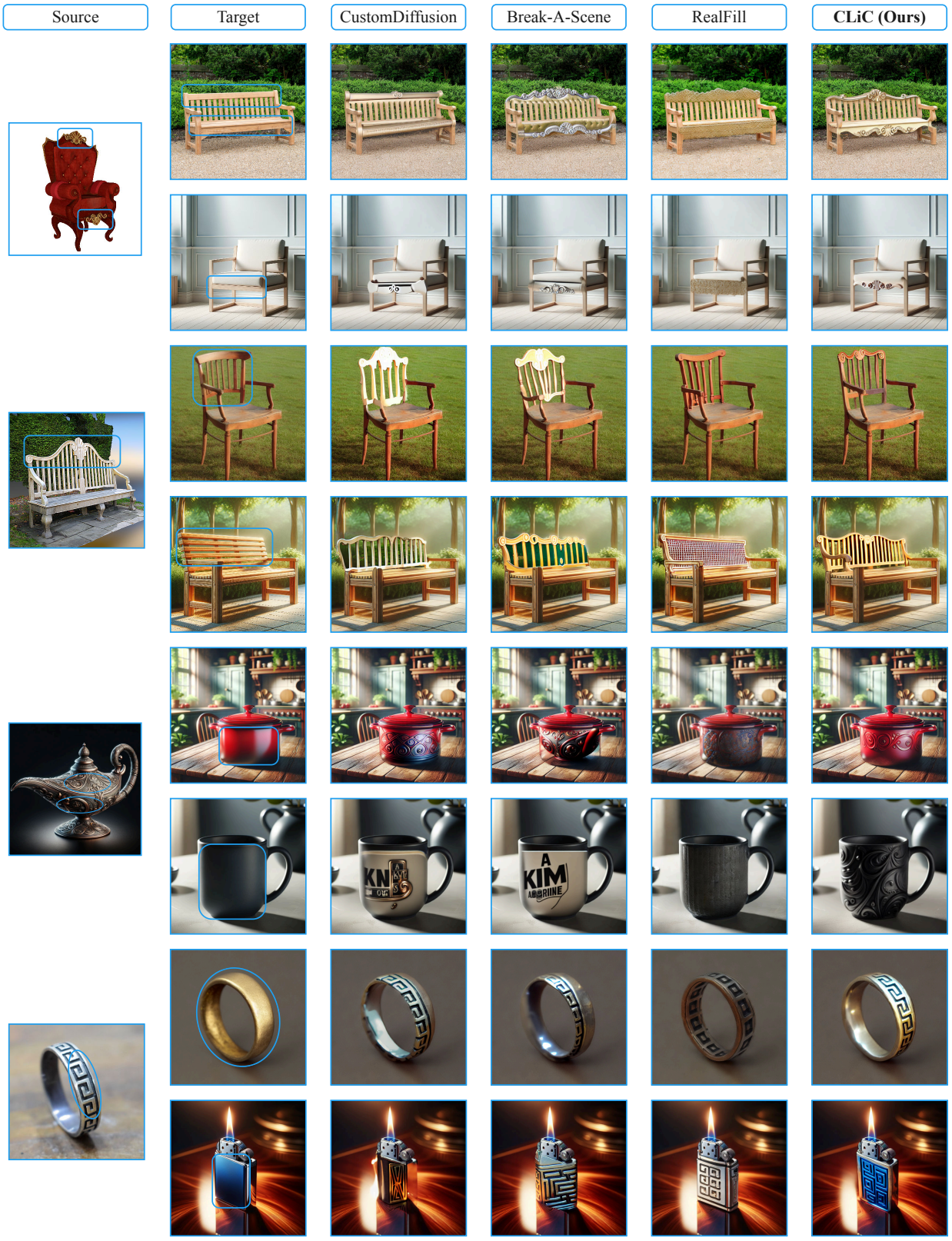

Figure 13. **Additional comparisons**. We further compare our concept transfer method with CustomDiffusion [22], Break-A-Scene [4], and RealFill [39].