

# Key-Locked Rank One Editing for Text-to-Image Personalization

YOAD TEWEL, NVIDIA, Tel-Aviv University, Israel

RINON GAL, NVIDIA, Tel-Aviv University, Israel

GAL CHECHIK, NVIDIA, Bar-Ilan University, Israel

YUVAL ATZMON, NVIDIA, Israel

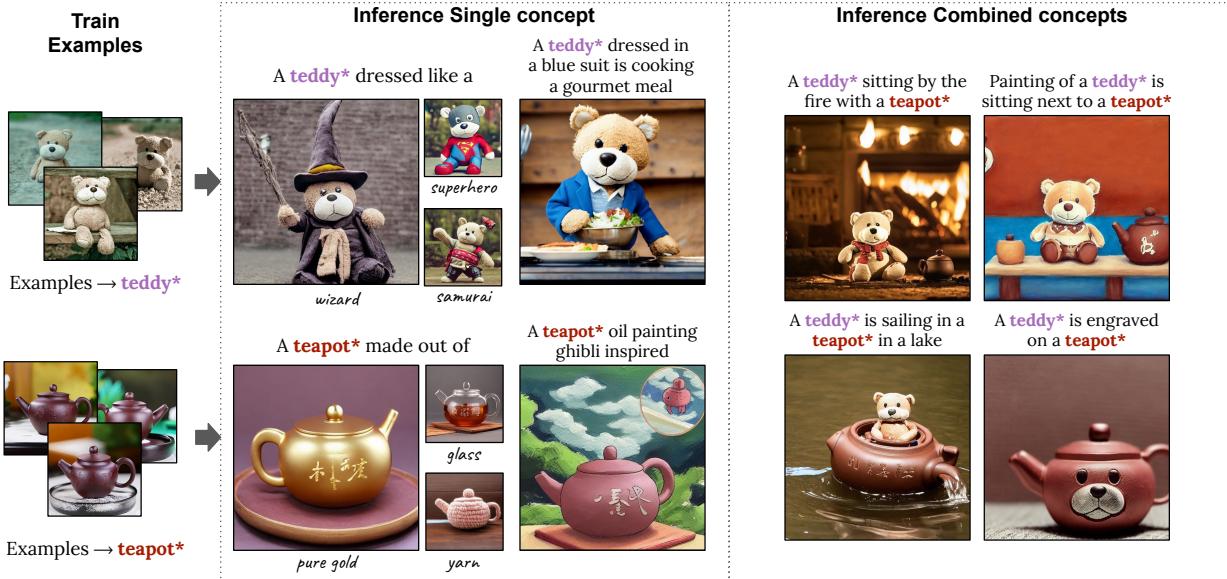


Fig. 1. (left) *Perfusion* enables image generation for personalized concepts with large changes in their appearance, pose, and context, using a compact model of only 100KB per concept, without compromising identity. (right) *Perfusion* can combine learned concepts at inference time, creating scenes which portray multiple concepts side-by-side, or even create interactions between them.

Text-to-image models (T2I) offer a new level of flexibility by allowing users to guide the creative process through natural language. However, personalizing these models to align with user-provided visual concepts remains a challenging problem. The task of T2I personalization poses multiple hard challenges, such as maintaining high visual fidelity while allowing creative control, combining multiple personalized concepts in a single image, and keeping a small model size.

We present *Perfusion*, a T2I personalization method that addresses these challenges using dynamic rank-1 updates to the underlying T2I model. *Perfusion* avoids overfitting by introducing a new mechanism that “locks” new concepts’ cross-attention Keys to their superordinate category. Additionally, we develop a *gated* rank-1 approach that enables us to control the influence of a learned concept during inference time and to combine multiple concepts. This allows runtime-efficient balancing of visual-fidelity and textual-alignment with a single 100KB trained model, which is five orders of magnitude smaller than the current state of the art. Moreover, it can span different operating points across the Pareto front without additional training.

Finally, we show that *Perfusion* outperforms strong baselines in both qualitative and quantitative terms. Importantly, key-locking leads to novel results compared to traditional approaches, allowing to portray personalized object interactions in unprecedented ways, even in one-shot settings.

Code will be available at our [project page](#)<sup>1</sup>.

## 1 INTRODUCTION

Text-to-image (T2I) personalization is the task of customizing a diffusion-based T2I model to reason over novel, user-provided visual concepts [Cohen et al. 2022; Gal et al. 2022; Kumari et al. 2022; Ruiz et al. 2022]. A user first provides a handful of image examples of the concept; then, they can then use free text to craft novel scenes containing these concepts. This workflow can be used in a wide range of downstream applications from virtual photo shoots through product design to generation of personalized virtual assets.

Current methods for personalization take one of two main approaches. They either represent a concept through a word embedding at the input of the text encoder [Cohen et al. 2022; Gal et al. 2022] or fine-tune the full weights of the diffusion-based denoiser module [Ruiz et al. 2022]. Unfortunately, these approaches are prone to different types of overfitting. As we show below, word embedding methods struggle to generalize to unseen text prompts. This is reflected in their textual-alignment scores which tend to be low. Fine-tuning methods can better generalize to new text prompts, but they still lack expressivity, as reflected in their textual and visual alignment scores which tend to be lower than our method. Moreover, tuning methods typically demand significant storage space, often in the range of hundreds of megabytes or even gigabytes. Lastly,

<sup>1</sup>Accepted to SIGGRAPH 2023

both approaches struggle to combine concepts that were trained individually, such as a teddy\* and a teapot\* (Fig. 1), in a single prompt.

Here we describe “*Perfusion*” a T2I personalization method aimed at answering all these challenges. It allows for expressive deformations of the concept while maintaining high concept-fidelity. It further enables inference-time combinations of concepts, and it has a small model size – roughly 100KB per concept. To achieve these goals, we focus on the cross-attention module of diffusion-based T2I models.

In typical diffusion-based T2I, an input text prompt is transformed into a sequence of encodings using a text encoder such as T5 [Raffel et al. 2020] or CLIP [Radford et al. 2021]. These encodings are then mapped to Keys and Values using learned projection matrices as part of the cross-attention module. Inspired by Balaji et al. [2022]; Hertz et al. [2022], we propose to view the effects of these projections as two different pathways: The Keys (K) are a “*Where*” pathway, which controls the layout of the attention maps, and through them the compositional structure of the generated image. The Values (V) are a “*What*” pathway, which controls the visual appearance of the image components.

Our main insight is that existing techniques fail when they overfit the *Where* pathway (Figures 2, 11), causing the attention on the novel words to leak beyond the visual scope of the concept itself. To address this shortcoming, we propose a novel “*Key-Locking*” mechanism, where the keys of a concept are fixated on the keys of the concept’s super-category. For example, the keys that represent a specific teddybear may be key-locked to the super-category of a teddy instead. Intuitively, this allows the new concept to inherit the super-category’s qualities and creative power (Figure 1). Personalization is then handled through the *What* (V) pathway, where we treat the Value projections as an extended latent-space and concurrently optimize them along with the input word embedding.

Finally, we describe how these components can be incorporated directly into the T2I model through the use of a *gated* rank-1 update to the weights of the K and V projection matrices. The gated aspect of this update allows us to combine multiple concepts at inference time by selectively applying the rank-1 update to only the necessary encodings. Moreover, the same gating mechanism provides a means for regulating the strength of learned concept, as expressed in the output images. This allows runtime efficient, inference-time trade-off of visual-fidelity with textual-alignment, without requiring specialized models for every new operating point. Empirically, *Perfusion* not only leads to more accurate personalization at a fraction of the model size, but it also enables the use of more complex prompts and the combination of *individually-learned* concepts at inference time.

In summary, our contributions are as follows: First, we investigate the overfitting observed in current personalization methods and propose a “*Key-Locking*” mechanism that mitigates it. Second, we propose a controllable rank-1 update mechanism for the network that achieves high object fidelity with only a 100KB footprint. Third, our approach efficiently spans the Pareto front of a single trained model to balance visual-fidelity and textual-alignment during runtime, while also being able to generalize to unseen operating

points. Finally, we demonstrate that *Perfusion* can outperform the state-of-the-art and enable object compositions at inference time.

## 2 RELATED WORK

**Diffusion based text-guided synthesis.** Recent advances in T2I generation have been led by pre-trained diffusion models [Ho et al. 2020] and particularly by large models [Balaji et al. 2022; Nichol et al. 2021; Ramesh et al. 2022; Rombach et al. 2021] trained on web-scale data [Schuhmann et al. 2021]. Our approach extends these pre-trained models to portray personalized concepts. It is applied with Stable-Diffusion [Rombach et al. 2021], but we expect that it can be applied to any T2I generator that uses a similar cross-attention mechanism [Saharia et al. 2022].

**T2I Personalization** The task of T2I personalization [Gal et al. 2022; Ruiz et al. 2022] aims to teach a generative T2I model to synthesize new images of a specific target concept, guided by free language. Current personalization methods either fine-tune the denoising network around a fixed embedding [Ruiz et al. 2022] or optimize a set of word embeddings to depict the concept [Agrawal et al. 2021; Cohen et al. 2022; Daras and Dimakis 2022; Gal et al. 2022]. [Kumari et al. 2022] is a concurrent work that fine-tunes the K and V cross-attention layers of the denoising network, along with a word embedding. It uses a closed-form optimization technique to combine concepts. In *Perfusion*, we lock the K pathway to the concept’s supercategory and use gated rank-1 editing instead of fine-tuning and subsequent optimization. This yields novel quality of results. We focused on K and V pathways independently of [Kumari et al. 2022].

**Rank-1 Model editing** In the field of natural language processing, significant effort was given to understanding and localizing the memory mechanisms of large language models. Specifically, it has been observed that the transformer feed-forward layers serve as key-value memory storage [Geva et al. 2022, 2020; Meng et al. 2022]. Recently, Bau et al. [2020]; Meng et al. [2022] introduced ROME, a rank-1 editing approach that updates these associative memory layers in order to modify the network’s factual knowledge. *Perfusion* seeks to apply similar ideas to T2I diffusion models. However, naïve rank-1 editing of diffusion models can lead to poor results, as also reported by [Kumari et al. 2022]. Our approach addresses the challenges of applying these methods to diffusion-based cross-attention layers. Moreover, *Perfusion* can combine multiple rank-1 edits using a dynamic gating mechanism rather than a static edit.

**Text-based image-editing.** The advent of powerful multi-modal models has brought with it an array of text-based editing methods [Bau et al. 2021; Gal et al. 2021; Patashnik et al. 2021]. With diffusion models, these range from single-image editing approaches [Brooks et al. 2022; Kawar et al. 2022; Meng et al. 2021; Mokady et al. 2022; Valevski et al. 2022; Wu et al. 2022; Zhang et al. 2022] to inpainting tasks [Avrahami et al. 2022; Nichol et al. 2021; Ramesh et al. 2022; Yang et al. 2022]. Most relevant to our work are the paint-with-words (PWW) approach introduced by Balaji et al. [2022], and prompt-to-prompt (P2P) [Hertz et al. 2022]. PWW biases the attention map toward a predefined mask during inference time. P2P edits a *given* generated image, by regenerating it with a new prompt while injecting the attention maps of the original image along the diffusion

process. In contrast to these methods, we do not edit *given* images but learn to represent a personalized concept that can be invoked in new prompts. Additionally, we do not override the attention map, but constrain the cross-attention Keys of the new concept. These are a contributing factor to the attention map, but they still allow for concept-specific modifications through the Query features.

### 3 PRELIMINARIES AND NOTATIONS

We begin with an overview of two mechanisms that our work leverages for personalization. The first is the cross-attention mechanism typically found in T2I diffusion models [Rombach et al. 2021]. The second is a recent approach for rank-1 editing of large language models [Meng et al. 2022].

#### 3.1 Cross-Attention in Text-to-Image models

In current T2I systems based on diffusion models (Fig. 3.A), an input text prompt is first converted into a sequence of word-embeddings. This sequence is then transformed into a sequence of encodings using a text encoder, such as CLIP [Radford et al. 2021]. Each encoding is then linearly projected through two cross-attention matrices:  $W_K$  and  $W_V$ . The results of these projections are known as “Keys” and “Values”. Formally, let  $M$  be the length of the input sequence,  $w \in \mathbb{R}^{M \times d_w}$  be a sequence of word-embeddings, each with dimension  $d_w$ , and  $e \in \mathbb{R}^{M \times d_e}$  be a sequence of encodings, each with dimension  $d_e$ . For each entry  $m \in M$  in the sequence, the encoding  $e_m \in \mathbb{R}^{d_e}$  is mapped by the two projection matrices into a “Key” vector  $K_m = W_K e_m \in \mathbb{R}^{d_k}$  and a “Value” vector  $V_m = W_V e_m$ . Concurrently, local image features are projected through a third matrix,  $W_Q$ , generating a spatial map of “Queries”. These are in turn projected onto the keys, yielding a per-encoding attention map:  $A_m = \text{softmax}\left(QK_m^T / \sqrt{d_k}\right)$  [Rombach et al. 2021]. Intuitively, this map informs the model about the relevance of the  $m^{\text{th}}$  word to each spatial region of the image. Finally, local image features are comprised by the “Values”, weighted by these maps:  $A \cdot V$  (Figure 3).

#### 3.2 Rank-1 Model Editing

Rank-one Model Editing (ROME) [Meng et al. 2022] is a recently introduced method for editing factual association in a pre-trained language model, such as GPT. ROME edits the weights of a single linear layer  $W$  in the network, so that given one *target-input*  $i_*$ , the layer will emit one *target-output*  $o_*$ <sup>2</sup>. To edit the model’s factual knowledge, ROME employs three steps, performed separately. (1) *Find the target-input  $i_*$  associated with the edited word (fact) in layer  $W$ .* ROME determines the target-input activations of the edited word in different prompts by passing them through the language model and averaging the activations at the word’s index. This gives the representation  $i_*$ . (2) *Find the target-output  $o_*$ ,* by optimizing the output activation of the layer for a specific goal — e.g. to modify the facts presented by the model’s final output.  $o_*$  is optimized over the output of a *single* word index. (3) *Update the layer  $W$  by solving a*

constrained least-squares problem, which has a closed-form solution

$$\hat{W} = W + \Lambda(C^{-1} i_*)^T. \quad (1)$$

Here,  $\Lambda = (\mathbf{o}_* - Wi_*) / [(i_*^T C^{-1})^T i_*]$ ,  $C$  is a constant positive definite matrix, that is pre-cached (Appendix C). The update in step 3 is performed after finding  $\mathbf{o}_*$ , thus affecting all sequence encodings, instead of just the single edited word as in step 2.

By limiting the update to a local, rank-1 change, ROME changes the information associated with a single fact without drastically altering the knowledge of the model. Our method leverages a similar mechanism to edit a text-to-image model and introduce new visual concepts.

### 4 METHOD

We aim to personalize a model in an expressive and efficient manner. A natural place to start then is by investigating the limitations of prior work in the field, and particularly *Textual Inversions* [Gal et al. 2022] and *DreamBooth* [Ruiz et al. 2022]. We notice that these methods, and Textual Inversion in particular, are susceptible to overfitting, where a learned concept becomes difficult to modify by changing the prompt that contains it. In Figure 2 we demonstrate that this issue originates in the attention mechanism, as the new concept draws attention beyond its visual scope. Additional examples showing this phenomenon are provided in Figure A.3.

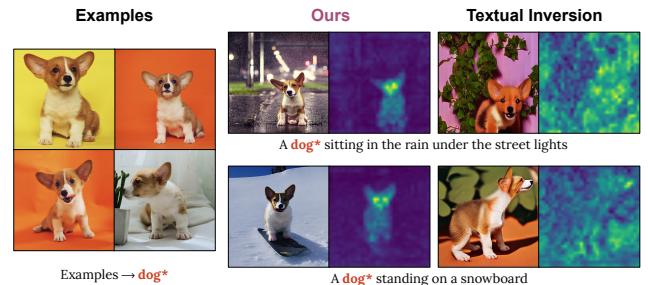


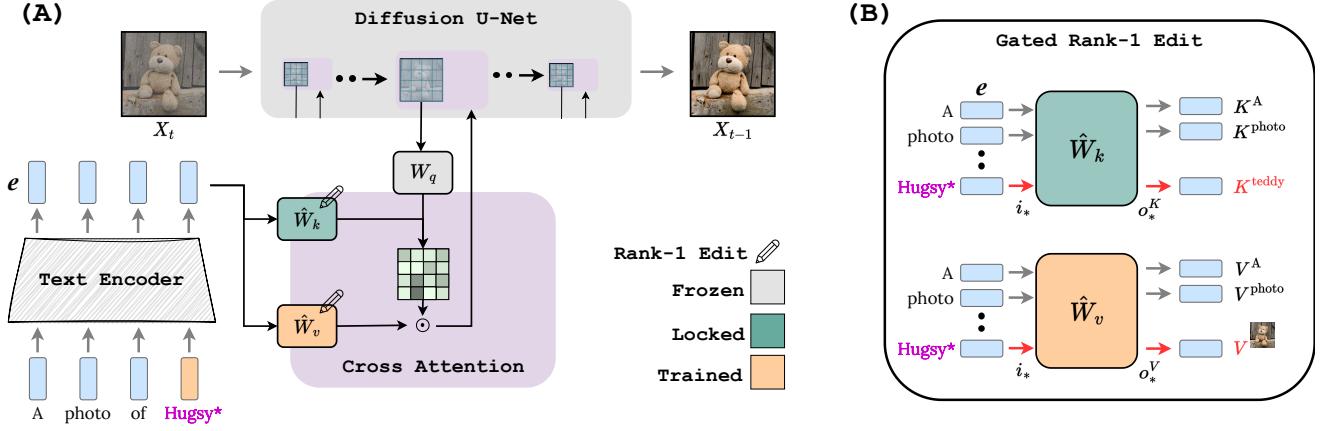
Fig. 2. **Attention overfit:** Typical overfit in Textual-Inversion (TI), caused by the attention of the learned embedding taking over the whole image. Here we visualize the attention maps that correspond to the “dog\*” word. The TI attention regions (right panel) are spread across the entire image rather than focusing on the object. This leads the generative process to ignore the rest of the prompt and depict only the “dog\*” concept.

Next, we describe **Perfusion**, an approach to overcome the problem through rank-1 layer editing. We outline a gating mechanism that provides better control at inference time and describe how to leverage it to compose concepts that were learned in isolation.

#### 4.1 Two conflicting goals and one Naïve Solution

Personalized T2I aims to achieve two goals: (1) *Avoid overfitting* to the example images, so the personalized concept can be generated in various poses, appearances, or context; and (2) *Preserve the identity* of the personalized concept in the generated image, despite being portrayed in a different pose appearance or context. There is a natural trade-off between these two goals. Methods that overfit the input examples tend to preserve identity, but then fail to match creative prompts that aim to place the concept in different contexts.

<sup>2</sup>Intuitively, the *target-input* plays a role of a key that is to be matched. To avoid confusion with transformers’  $K$  and  $V$  pathways, we use “target-input” and “target-output” instead of “key” and “value” from [Meng et al. 2022].



**Fig. 3. Architecture outline (A):** A prompt is transformed into a sequence of encodings. Each encoding is fed to a set of cross-attention modules (purple blocks) of a diffusion U-Net denoiser. Zoomed-in purple module shows how the Key and Value pathways are conditioned on the text encoding. The Key drives the attention map, which then modulates the Value pathway. **Gated Rank-1 Edit (B):** *Top:* The K pathway is locked so any encoding of  $e_{\text{Hugsy}}$  that reaches  $\hat{W}_k$  is mapped to the key of the super-category  $K^{\text{teddy}}$ . *Bottom:* Any encoding of  $e_{\text{Hugsy}}$  that reaches  $\hat{W}_v$ , is mapped to  $V^{\text{Hugsy}}$ , which is learned. The gated aspect of this update allows to selectively apply it to only the necessary encodings and provides means for regulating the strength of learned concept, as expressed in the output images.

**The Where Pathway and the What Pathway.** To improve both of these goals simultaneously, our key insight is that models need to disentangle what is generated from where it is generated. To this end, we leverage the interpretation of the cross-attention mechanism described in section 3.1. The  $K$  pathway – the one associated with the “Keys”, is related to creating the attention map. It thus serves as a pathway for controlling *where* objects are located in the final image. In contrast, the  $V$  pathway is responsible for the features added to each region. In this sense, it can control *what* appears in the final image. We therefore interpret  $K$  mappings as a “*Where*” pathway and  $V$  mappings as a “*What*” pathway, and this interpretation guides our proposed method:

**4.1.1 Avoid overfitting.** In preliminary experimentation, we noticed that when learning personalized concepts from a limited number of examples, the model weights of the *Where* pathway ( $W^K$ ) are prone to overfit to the image layout seen in these examples. Figure 2 illustrates this problem showing that the personalized examples may ‘dominate’ the entire attention map, and prevent other words from affecting the synthesized image. We thus aim to prevent this attention-based overfitting by restricting the *Where* pathway.

**4.1.2 Preserving Identity.** In Image2StyleGAN, Abdal et al. [2019] proposed a hierarchical latent-representation to capture identities more effectively. There, instead of predicting a single latent code at the generator’s input space, they predicted a different code for each resolution in the synthesis process. We propose the *What* ( $V$ ) pathway activations as a similar latent space, given their compact nature and the multi-resolution structure of the underlying U-Net denoiser.

**4.1.3 A Naïve Solution.** To meet both goals, consider this simple solution: Whenever the encoding contains the target concept, ensure that its cross-attention *keys* match those of its supercategory, which

we call **Key Locking**. Additionally, we want the cross-attention *values* to represent the concept in the multi-resolution latent space. For example, as illustrated in Figure 3 (right-top), given image examples of our teddy bear named Hugsy, when the encoding includes “Hugsy” the  $V$  projection emits a concept-specific code, while the  $K$  projection is targeted to emit keys for the super category  $K^{\text{teddy}}$ .

One way to implement this idea would be a simple vector replacement - simply swapping out the keys and values assigned to the encoding at the personalized concept’s index. However, this fails to account for the cross-word information sharing in the text encoder. By the time the encoding reaches the denoiser’s cross-attention layers, its features are already influenced by the features of other words in the text, and in turn, influence them as well. We want to ensure that our implementation accounts for this influence, and correctly modifies the Key and Value activations for any such influenced words.

A natural solution is then to edit the weights of the cross-attention layers,  $W_V$  and  $W_K$  using ROME. Specifically, when given a target-input  $i_{\text{Hugsy}}$  we enforce the  $K$  activation to emit a specific target-output  $o_{\text{Hugsy}}^K = K^{\text{teddy}}$ . Similarly, given a target-input  $i_{\text{Hugsy}}$ , we enforce the  $V$  activation to emit a learned output  $o_{\text{Hugsy}}^V = V^{\text{Hugsy}}$  see Figure 3 (right-bottom). Now, for any word, if its encoding contains a component parallel (aligned) to  $i_*$ , then their activation outputs will also be modified accordingly.

Unfortunately, applying ROME to this task faces two challenges: **Challenge 1:** Training with ROME leads to a mismatch between training and inference. This is because during training in “step 2”, ROME optimizes only the target-output  $o_*$  associated with one specific entry in the prompt ( $m^{\text{th}}$ -index). However, as noted above, when performing the rank-1 matrix update in “step 3” the change is expected to affect the projections of other words in the prompt. Indeed, we have observed that this results in a train-test mismatch

that substantially degrades the fidelity of the reconstructed concept. **Challenge 2:** A similar effect also prevents us from combining more than one learned concept, as their effects on the projections are not well-disentangled. Moreover, these new concepts are associated with multiple target-inputs  $i_*$ , which may themselves be inherently entangled (e.g. if the concepts share related semantics). Together, these lead to the creation of visual artifacts when attempting to combine concepts at inference-time.

To address these challenges we propose to align the training and inference steps of ROME, and introduce a new gating mechanism. Both components are described below.

#### 4.2 Gated Rank-1 Model Editing for Personalized T2I

**Training end-to-end to address train-test mismatch.** To address the first challenge, ROME’s mismatch between training and inference, we propose to unify the second and third steps of ROME. As such, the target-output optimization and matrix update occur together during training. The network learns to account for any effects on other prompt-parts, avoiding the train-inference mismatch.

To do so, we rewrite the weight update of ROME, to characterize the output  $h$  of layer  $\hat{W}$  when presented with an input  $e_m$ . This yields

$$h = We_m^\perp + o_* \text{sim}(i_*, e_m) / ||i_*||_{C^{-1}}^2, \quad (2)$$

Here,  $\text{sim}(i_*, e_m) := i_*^T (C^{-1})^T e_m$  measures the similarity of  $e_m$  with  $i_*$  in a metric space defined by  $C^{-1}$  [Atzmon et al. 2015],  $||i_*||_{C^{-1}}^2 := \text{sim}(i_*, i_*)$  measures the energy of  $i_*$  in the same metric space, and  $e_m^\perp := e_m - i_* \text{sim}(i_*, e_m) / ||i_*||_{C^{-1}}^2$  is the component of  $e_m$  that is orthogonal to  $i_*$  in the metric space. Intuitively, the right additive term in Eq. (2) maps the  $i_*$  component of the word encoding ( $e_m$ ) to  $o_*$ . The left term nulls the  $i_*$  component from the word encoding and maps the remaining using the pretrained matrix  $W$ . We provide more details in Appendix A and B.1.

Given this characterization, we replace the forward pass of each layer by updating it using Eq. (2) as the layer’s forward pass. This ensures that the same update expression is used for both training and inference, eliminating the mismatch. Combined with an online estimation of  $i_*$  (Section 4.4), it enables end-to-end training with ROME, rather than individually applying its 3 steps.

**Using gated rank-1 update for combining concepts.** Combining individually learned concepts at inference time is a hard challenge. In initial experiments, we tested adding concepts one by one, by editing  $W$  using  $\hat{W} = W + \sum_{j=1..J} \Lambda^j (C^{-1} i_*^j)^T$  as a variation of equation Eq. (1). We found this approach introduced visual artifacts, even when the prompt only included a single learned concept. We hypothesize that this problem arises because the different  $i_*$ s of individually learned concepts may interfere with each other. For example, they may have related representations if the concepts share semantics. Ideally, a concept update occur only when input encoding has sufficient energy regarding the concept, and attenuated otherwise. By doing so, we can ensure that the model update is only applied to the relevant concept, and not to others.

To address this challenge, we use a *gating* mechanism to selectively allow or attenuate the influence of each concept on the layer output. Here, we note that the update rule of Eq. (2) already includes

a *linear* gating mechanism  $\text{sim}(e_m, i_*) / ||i_*||_{C^{-1}}^2$ , which is close to 1 when  $e_m = i_*$ . However, lower similarity values may not be sufficiently attenuated. We therefore propose to increase the influence of  $\text{sim}$  by wrapping the  $\text{sim}(e_m, i_*) / ||i_*||_{C^{-1}}^2$  value with a sigmoid function, which has hyper-parameters for bias and temperature. This way, the weight updates are sharply concentrated on inputs that strongly correspond to the personalized concept.

Therefore, the forward pass of each layer update during both training and inference time of a single concept, is

$$h = We_m^\perp + o_* \sigma \left( \frac{\text{sim}(i_*, e_m) / ||i_*||_{C^{-1}}^2 - \beta}{\tau} \right) \quad (3)$$

where,  $\sigma$  is a sigmoid activation function,  $\tau$  is the temperature and  $\beta$  a bias term. This implementation ensures that weight updates are only applied to encoding components that align (parallel) with  $i_*$ , i.e. those belonging to the new concept, or influenced by it.

This non-linear gating mechanism therefore provides **two important benefits**: First, it allows us to better separate the influence of individually learned concepts during inference time. Second, even for a single concept, it allows for *inference-time* control over the influence of the concept. By adjusting the values of the sigmoid hyper-parameters, the bias and the temperature, we can trade visual fidelity with textual alignment and vice-versa.

In the next section we expand on how to generalize this formulation to combine multiple concepts.

#### 4.3 Inference

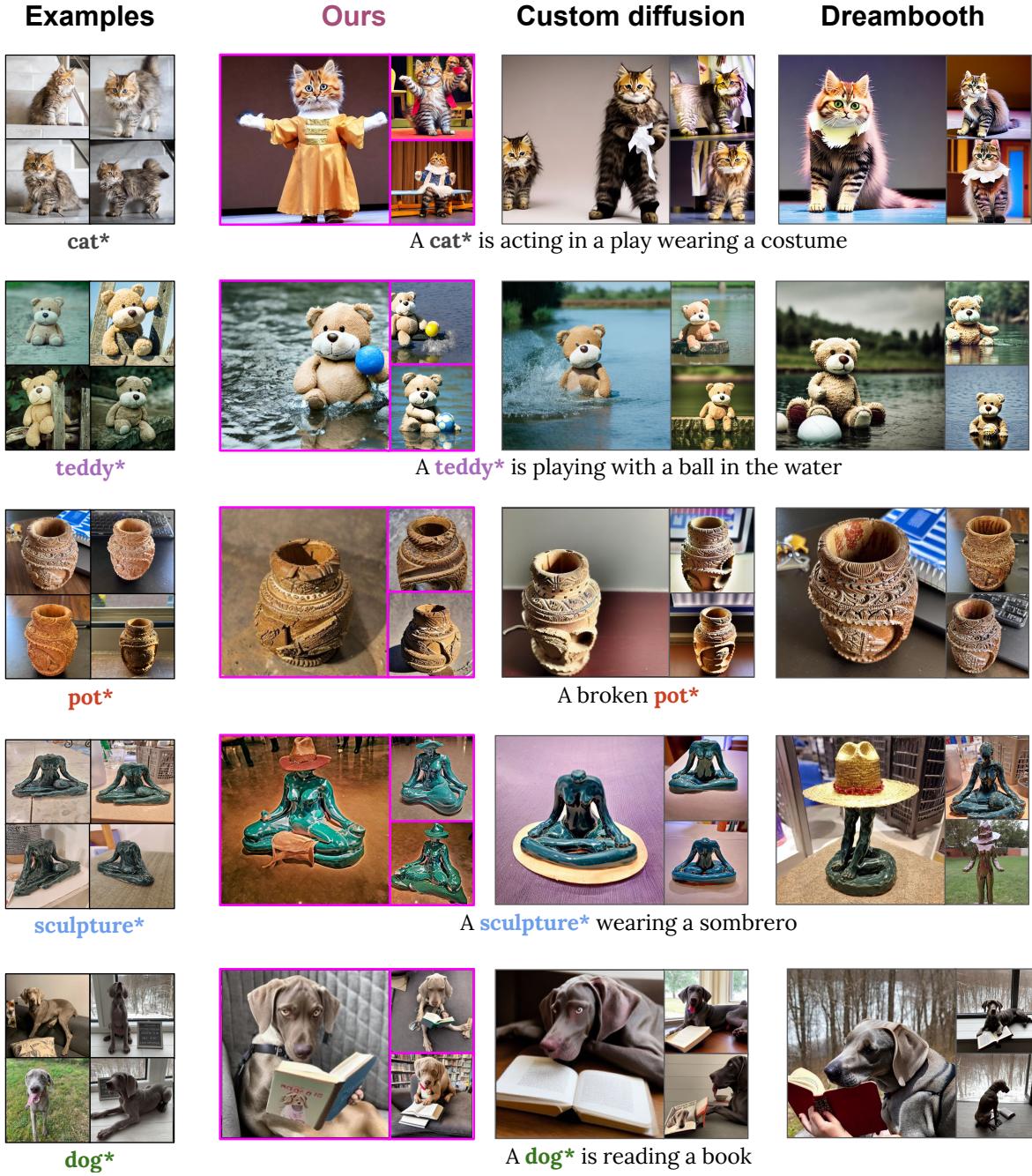
**Single concept:** For inference with a single trained concept, we simply apply Eq. (3) to the forward pass of each edited cross-attention layer. We can control the strength of the depicted concept by changing the values of the sigmoid’s  $\tau$  and  $\beta$  at inference time.

**Combining multiple concepts:** To combine concepts that were trained in isolation, we extend equation Eq. (3) to include multiple concepts  $\{i_*^j, o_*^j\}_{j=1..J}$ . For that, we first generalize  $e_m^\perp$  to be orthogonal to the sub-space spanned by all the  $\{i_*^j\}_{j=1..J}$  in the metric space, which we denote as  $e_m^{\perp J}$ . For the right term we simply sum the gated responses from all the concepts. The final expression is:

$$h = We_m^{\perp J} + \sum_{j \in 1..J} o_*^j \sigma \left( \frac{\text{sim}(i_*^j, e_m) / ||i_*^j||^2 - \beta_j}{\tau} \right), \quad (4)$$

where  $e_m^{\perp J} = e_m - \sum_{j \in 1..J} u_j \text{sim}(u_j, e_m)$ , and  $u_j$  relates to a basis vector in the metric space, after being projected back to the text encoder space by an inverse Cholesky root  $L^{T^{-1}}$ . The derivation of  $e_m^{\perp J}$  is provided in Appendix B.

**4.3.1 Global Key-Locking:** Key-Locking ensures that a concept’s Key is correctly aligned with its superconcept. However, it does not ensure that the text-encoder handles the concept in the same way it would have handled the superconcept and its correlations to the other words in the encoding. We also investigate an inference time method to align Key-locked concepts to an entire prompt. We refer to this variant as *global key-locking*, and to our vanilla mechanism as *local key locking*. We describe the details in Appendix C.



**Fig. 4. Generation results with single concept examples.** For each concept, we show exemplars from our training set, along with generated images, their conditioning texts and comparisons to Custom-Diffusion (CD) and Dreambooth (DB) baselines. *Perfusion* can enable more animate results, with better prompt-matching and less susceptibility to background traits from the original image. Note in particular the improved garments and theatrics on our cat (top), or the prompt-appropriate gaze and posture when instructing our dog to read a book (bottom). For some prompts, the baselines simply copy the content from the training set (*e.g.* the pot).

#### 4.4 Implementation details

**Online estimation of  $i_*$ :** We use the following exponential moving average expression to estimate  $i_*$  during training time:  $i_* := 0.99i_* +$

$0.01e_{\text{concept}}$  where  $e_{\text{concept}}$  corresponds to encoding of the concept word at the output of the text encoder.



Fig. 5. **Additional generation results with multi concept examples.** We show pairs of concepts interacting, and compare to CD. Except for the *teddy\** prompt, all prompts are from CD paper and use the images provided by the paper. In the *teddy\** example, *Perfusion* portrays it with the *sunglasses\**, while CD omits the *sunglasses\**. In the watercolor painting *Perfusion* better preserve the chair shape. In the *table\** example, *Perfusion* better preserve the table color.

**Pseudo Code:** Appendix D provides pseudo code for the rank-1 editing module of *Perfusion*.

**Zero-Shot Weighting Loss:** Training with few image examples is prone to learning spurious correlations from the image background. To decorrelate the concept from its background we weigh the standard conditional diffusion loss by a soft segmentation mask attained from a zero-shot image segmentation model [Lüdecke and Ecker 2021]. Mask values are normalized by their maximum value.

**Applying *Perfusion* to multiple layers:** Similar to [Gal et al. 2022], for each concept we choose a single word for a supercategory name. We use that word to initialize its word embeddings and treat the embeddings as learned parameters. We apply *Perfusion* editing to all cross-attention layers of the UNet denoiser. For each of the K pathway layers ( $l$ ), we precompute and freeze the  $\mathbf{o}_*^{K:l}$  to be  $\mathbf{o}_*^{K:l} = \mathbf{W}_K^l \mathbf{e}_{\text{superclass}}$  with a prompt saying “A photo of a <superclass\\_word>”, and we update  $\mathbf{i}_*$  as training progresses. On each of the V pathway layers, we treat  $\mathbf{o}_*^{V:l}$  as learned parameters.

**Training details:** We train  $\mathbf{o}_*$  with a learning rate of 0.03, for the embedding we set a learning rate of 0.006. We use a batch size of 16 using Flash-Attention [Dao et al. 2022; Leflaudeux et al. 2022]. We only use flip augmentations  $p = 50\%$ . We do not flip asymmetric objects. We use a validation set of 8 prompts, sampled every 25 training steps, and select the step with the model that maximizes the harmonic mean between a CLIP image similarity score, and a CLIP text similarity score. We describe the CLIP metrics in more detail in the experimental details. To condition the generation, we randomly sample neutral context texts, derived from the CLIP ImageNet templates [Radford et al. 2021]. The list of templates is provided in the supplementary materials.

Our approach is trained on a single A100 GPU for an average of 210 steps, taking ~4 minutes, and a maximum of 400 steps (~7

minutes). This training requires  $\times 2 - \times 3$  less compute compared to concurrent work [Kumari et al. 2022] and utilizes 27GB of RAM.

**Sigmoid hyper-parameters:** At training time, we set the sigmoid bias and temperature to  $b = 0.75, T = 0.1$ . At inference time we typically use a temperature of 0.15 and bias of 0.6 – 0.75 for local key lock, or 0.4 – 0.6 for global key lock.

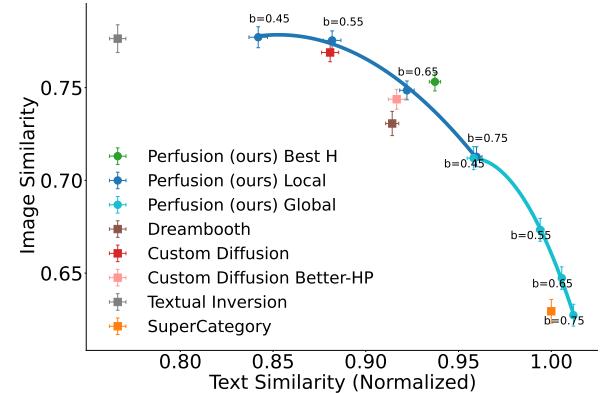


Fig. 6. **Visual - Textual Similarity Plane:** With just a single 100KB trained model and run-time parameter choices, *Perfusion* (blue and cyan) spans the Pareto front. Error bars denote 95% confidence intervals.

## 5 EXPERIMENTS

We demonstrate that *Perfusion* outperforms strong baselines. We conduct both a qualitative comparison and a quantitative evaluation, demonstrating that it spans the visual-fidelity and text-alignment Pareto front and achieves higher fidelity results with more complex prompts, despite using only a fraction of the parameters. Then,

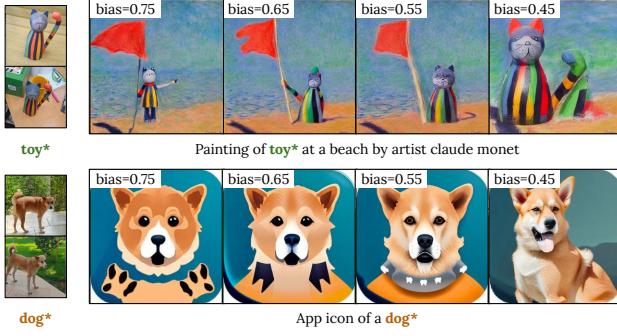


Fig. 7. **Balancing visual-fidelity and textual-alignment:** Controlling the bias threshold of the sigmoid allows to balance the trade-off between visual-fidelity and the textual-alignment. A high bias value reduces the concept’s effect, while a low bias value makes it more influential.



Fig. 8. **1-shot training results:** When training with a single image, our method can generate images with both high visual-fidelity and textual-alignment. We provide quantitative results in Figure 9(a).

we study the properties of *Perfusion* through an ablation study (Section 5.3).

**Compared Methods:** (1) **Perfusion:** Our method as described in Section 4. We use a *single* trained model for each concept, but we show results spanning different sigmoid biases and with *local* and *global* locking. These parameter adjustments are all applied during runtime. (2) **Perfusion Best H:** For each class, we automatically choose the run-time variant with the best harmonic mean over text and visual similarities (see metrics). (3) **DreamBooth** [Ruiz et al. 2022]: A SoTA approach that fine-tunes all parameters of the denoiser’s U-Net. We use the implementation of Patil and Cuencia [2022]. (4) **Textual-Inversion (TI)** [Gal et al. 2022]: A SoTA approach that only optimizes word embeddings. We use the Stable Diffusion implementation from the official repository, with the parameters the authors report for LDM [Rombach et al. 2021]. (5) **Custom-Diffusion (CD):** A concurrent work that trains the K and V cross-attention pathways, and also the word-embedding. We use the official implementation and hyperparameters. (6) **Custom-Diffusion Better-HP:** CD trained with 200 steps, which we found to improve the text similarity score. (7) **SuperCategory:** A text only baseline; We replace the concept word with its super class. All methods were applied to a pre-trained Stable Diffusion checkpoint *v1.5* [Rombach et al. 2021].

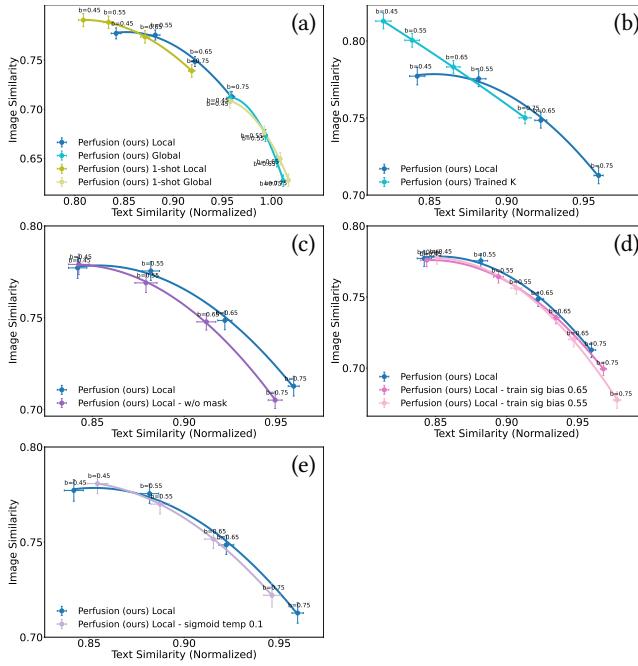
**Dataset:** *Concepts:* For fair and unbiased evaluation we used concepts from previous papers: 6 concepts from CD, 2 from TI, and 3 from or similar to DB, for a total of 11 personalization concepts. These are from four groups: 4 toys/figurines and 3 pets (grouped as “animated”), 2 containers, 1 furniture, and 1 wearable accessory. *Prompts:* We use two types of prompts. First, 19 prompts that were shared across all concepts. These only change the scene, but do not deform the concept. We name them “*shape-preserving*” prompts. Second, *per-group* prompts. These are more specific to the group the concept belongs, and often induce a deformation to the concept appearance, like “A *broken pot*”, or “A *cat* is *acting* in a play *wearing a costume*”. In total, we have 86 unique prompts, with an average of 43 prompts per class. Out of the 86 unique prompts, 42 were randomly selected from those used by Custom-Diffusion.

**Evaluation Metrics:** Like TI, we report the results on a 2D plane to illustrate the balance between visual and text similarity. But unlike TI, which only uses prompts like “a photo of a *S\**” to evaluate image similarity, we use all shape-preserving prompts to also measure concept fidelity in new scenes..

We compute the following evaluation metrics: (1) **Image similarity** is the average pairwise CLIP cosine-similarity between the concept images and the generated images from the shape-preserving prompts. (2) **Text Similarity** is the average CLIP similarity between all generated images and their textual prompts, where we omit the placeholder *S\** (i.e. “A is dressed like a wizard”). To calibrate between different prompts, we normalize text scores by comparing them to scores of images generated with a supercategory word instead of the learned word. All similarity scores are balanced “per-class” [Samuel et al. 2020]. Namely, we compute the mean score per concept class, and then average all class scores. For each concept, we sample 8 images per prompt, by 50 DDIM steps and a guidance scale of 6.

## 5.1 Results

In Figure 6, we illustrate the results on a plane that shows the trade-off between visual and textual similarity. *Dreambooth* successfully balances compromise both visual and textual fidelity. However, it obtains lower scores on both metrics when compared to *Perfusion* and CD. *Textual-Inversion* struggles to generalize to new prompts, showing low textual-similarity score. *Custom-Diffusion* tends to favor visual similarity, even at the cost of overfitting to the target. We observe that with a small hyper-parameter change, *Custom-Diffusion Better-HP*, can instead be balanced toward textual similarity. *Perfusion* outperform these baselines and pushes forward the pareto front. Notably, we can span this front using inference-time parameter modifications, which allows a user to control this trade-off based on their desired qualities. In practice, this allows us to easily select the best operating point for each class, leading even greater performance (*Perfusion Best H*). Note that *Perfusion* achieves these gains while requiring only 100KB of per-concept parameters, compared to several GBs in DB and nearly 100MB in CD. **Importantly**, the comparison reveals that Key-Locking significantly improves textual similarity without significant harm to the visual similarity.



**Fig. 9. Ablation study:** We show the Pareto front for various ablation conditions. (a) **1-shot training:** our method trained with a single image per concept. (b) **Trained K:** We train the key pathway as well, instead of locking it. (c) **Usage of Zero-shot Mask During Training:** We compare *Perfusion* with and without the zero-shot mask. (d) **Sigmoid Bias During Training:** We compare training with different values of sigmoid bias (our default value is 0.75). (e) **Sigmoid Temperature During Inference:** We compare inference with sigmoid temperature of 0.1 to the default temperature of 0.15.

## 5.2 User Study

We further evaluate the models through two user studies conducted with Amazon Mechanical Turk. In the first study, raters were given two images of a concept and a prompt. They were asked to rank images generated by the three methods (*Perfusion*, CD, DB), based on how well they portrayed the concept according to the prompt. We used 11 concepts, 24 prompts per concept, with 8 responses per prompt. *Perfusion* was selected first with an average rank of  $2.18 \pm 0.02$  (SEM), CD was 2nd ( $2.06 \pm 0.02$ ) and DB was last ( $1.75 \pm 0.02$ ), demonstrating a preference for our approach. For the second study, we investigated whether *Perfusion* harms the generative prior. To do so, we compared *Perfusion* to “vanilla” stable-diffusion (SD). Raters were shown a prompt and two generated images, one by *Perfusion* and another by SD. They were asked to rate the images according to their realism, using a score between 1 to 3 (best). We gathered the same number of concepts, prompts and responses as the first study. *Perfusion* had an average score of  $1.885 \pm 0.017$  (SEM), while SD had a score of  $1.894 \pm 0.017$ . The results are statistically indistinguishable, demonstrating that *Perfusion* can preserve the generative prior. See Appendix I for additional details on these experiments.

## 5.3 Ablation study

In Figure 9 we study in greater depth the properties of *Perfusion* by an ablation study. We show the trade-off between visual and textual similarity for the following conditions.

- (a) **1-shot:** We compare between training our method with all the training examples (average of 6.5 image for each concept), to training with just a single example for each concept. We observe that training with just a single example introduces slight overfitting.
- (b) **Key-locking:** We compare between our method with key-locking to our method with trained key projection layers. It is evident that key-locking shifts the Pareto curve to the right - meaning less overfit. This result confirms our hypothesis that locking the key projection layers leads to better textual-alignment and enables complex deformations of the learned concept.
- (c) **Zero-Shot Masking Loss:** We compare the effects of training with and without a zero-shot mask. We notice that using zero-shot mask tends to improve the textual similarity, which mean it helps reduce the overfitting.
- (d) **Sigmoid Train Bias:** We compare between different values of the Sigmoid biases used during training time. We notice that using a higher bias results in better Pareto front.
- (e) **Sigmoid Inference Temperature** We compare between different values of the Sigmoid temperature used during inference time. We notice that using inference-time Sigmoid temperature that is higher than the train-time Sigmoid temperature results a better pareto front. Generally temperature of 0.15 tends to work better.

## 6 QUALITATIVE VISUAL COMPARISONS

Next, we provide qualitative comparisons that reveals the strengths of our approach, along with examples of our main failure mode.

**Single Concept Text-guided Synthesis:** Figures 1, 4 and 15 show our ability to compose novel scenes when using *Perfusion* and compare them with strong baselines. For each concept, we show exemplars from our training set, along with generated images and their conditioning texts. Our approach allows making deformations to the concept appearance without losing its identity. At the same time, it correctly encapsulates the semantic qualities of both the concept and the prompt. For example, notice how we can fully customize the teddy\* concept in Figure 1 and the cat\* in Figure 4 with different garments, without compromising their identities, and at the same time allow them to interact with the scene. We can also change the material of the teapot\* to pure gold or transparent glass, while retaining its distinctiveness. In Fig. 4 *Perfusion* is the only one that can shatter the pot\* concept. Notice how it can change the dog\* posture, making it appear as though it is engaged in reading a book, with its eyes focused on the text and its paws grasping the book. In Fig. A.2 we provide additional results including Textual Inversion.

**Multi Concept Text-guided Synthesis:** Figures 1 and 5 show our ability to compose novel scenes with multiple concepts, when using *Perfusion* and compare them with CD’s “optimization” approach to combine individual concepts. We use their provided images when comparing with prompts from their paper, otherwise



Fig. 10. **Comparing lock types:** *Global key-locking* allows for more visual variability and can accurately portray the nuances of an object or activity, like when depicting the cat *in a human-like posture* reading a book or wearing a chef outfit. *Local key-locking* also has successes, but they are not as effective as global key-locking. Finally, *Trained-K* has better compatibility with the training images, but it sacrifices its alignment with the text.

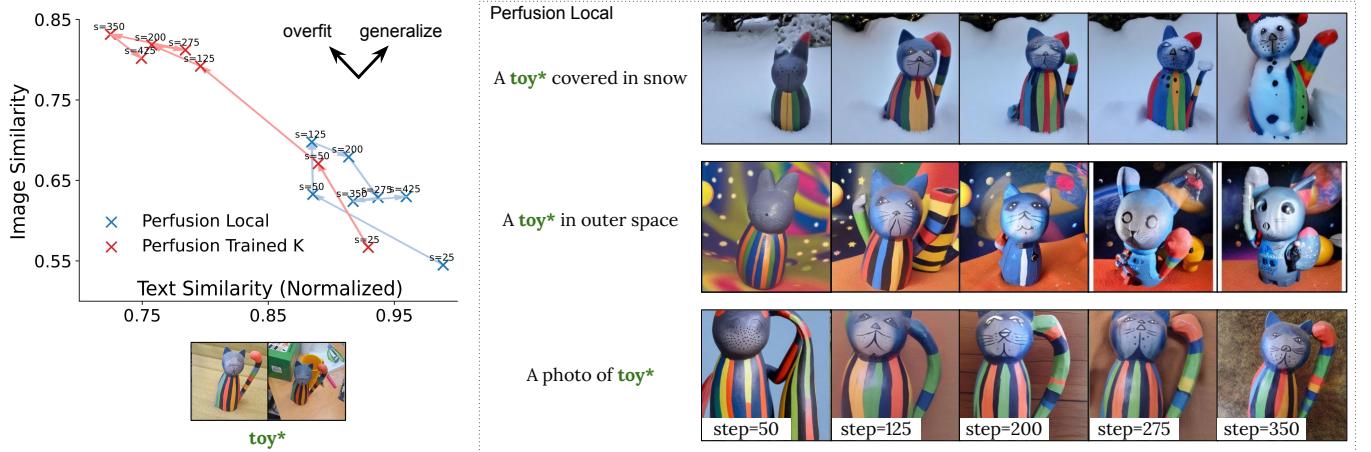


Fig. 11. **The impact of Key-Locking on training dynamics:** Comparing *Trained K* (without lock) and *Local Key-Locked* on the plane of fidelity vs. textual alignment, where ‘s’ denotes the training step. *Trained K* overfits the training images, but negatively affects textual alignment. Conversely, *Local Key-Locked* training exhibits “over-generalization”, leading to improved textual alignment, but reduced visual fidelity. As training progresses, qualitative examples reveal that the model becomes increasingly toy-like. This suggest that the V features learn latent properties of the supercategory to improve the alignment between the Q and K features. Finally, the cat-toy in this example demonstrates a failure case as the V features pick up the toy supercategory too early in the training.

with use “Better-HP”. Notice how *Perfusion* can compose the teddy\* and teapot\* in different scenes, or how it allows the teapot\* to hold the teddy\* while sailing. When comparing with CD, we observe similar or better results. For example, observe how the water-color painting better preserve the chair identity, or how the teddy\* can wear the sunglasses\* successfully.

**Balancing visual-fidelity and textual-alignment:** Figure 7 provides qualitative examples for balancing the visual-fidelity versus the textual-alignment, by adjusting the sigmoid bias threshold. Higher bias values reduce the impact of the concept, while lower values give it more prominence in the generated image. This is because the concept energy is spread across multiple encodings in the

text encoder, not just the one corresponding to the concept word. Lowering the bias increases its influence on all relevant encodings.

Next, we demonstrate several aspects of the key-locking mechanism. We start by comparing global key-locking local key-locking and no key-locking. Then, we show what happens when we lock concepts to different super-categories. Finally, we study the training dynamics of local key-locking and show an “over-generalization” phenomena that makes it over-align with the supercategory.

## 6.1 Key Design Decisions

Next, we demonstrate and discuss key design decisions. We start by comparing global key-locking local key-locking and no key-locking. Then, we show what happens when using vanilla ROME, when there a train-inference mismatch, and the weight update is performed only after the optimization step.

**Comparing lock types:** Fig. 10 compares *global key-locking*, *local key-locking* and *trained-K* (no key-locking, K pathway is trained like the V pathway). We find that global locking allows to generate rich scenes, portray better the nuances of the object attributes or activities, and in general allow more visual variability of the concept, compared to local key locking and trained-K. For example notice the cat depicted in human-like postures while reading a book, or when wearing a chef outfit. Local locking also has some successes but they are weaker. Finally, trained-K is more aligned with the postures and appearance of the training images, while sacrificing alignment with the text.

**Train-inference mismatch, when using vanilla ROME:** Fig. A.1 illustrates the mismatch between generated images during training and inference when editing with vanilla ROME. This results in corrupted images during inference.

## 6.2 Robustness Analysis

Next, we will show how our approach performs in different situations, highlighting both its strengths and weaknesses.

**One-shot learning:** We compare training our method with all examples (average of 6.5 images/concept) to using only one example per concept. In Figures 8 and 9(a), we note a slight overfit when training with only one example.

**Zero-shot transfer to fine-tuned checkpoints:** A Perfusion concept trained using a vanilla diffusion-model could generalize to fine-tuned variants of the model. Fig. 12 shows transfer abilities to two popular variants of Stable-Diffusion: InkPunk-v2 [Envvi 2022] and Protagen-v3.4 [darkstorm2150 2022].

**Uncurated samples:** Fig. 14 shows that a batch size of 8 is typically sufficient to ensure several good samples.

**Locking to unusual super-categories** Figure 13 shows how the concepts are portrayed when locked to unusual super-categories. We observe that the concepts “inherit” the qualitative outline of the unusual super-category. When the pot is locked to a shoe or a clock, it is portrayed in the outline of a shoe or clock, but in the style of the pot. When the cat is locked to a lamp, it becomes illuminated.

**The impact of Key-Locking on training dynamics:** The left panel of Figure 11 shows the comparison between *Trained K* and *Local* Key-Locked training for various training steps (“s”). As training progresses, *Trained K* (no-lock) overfits the training images,

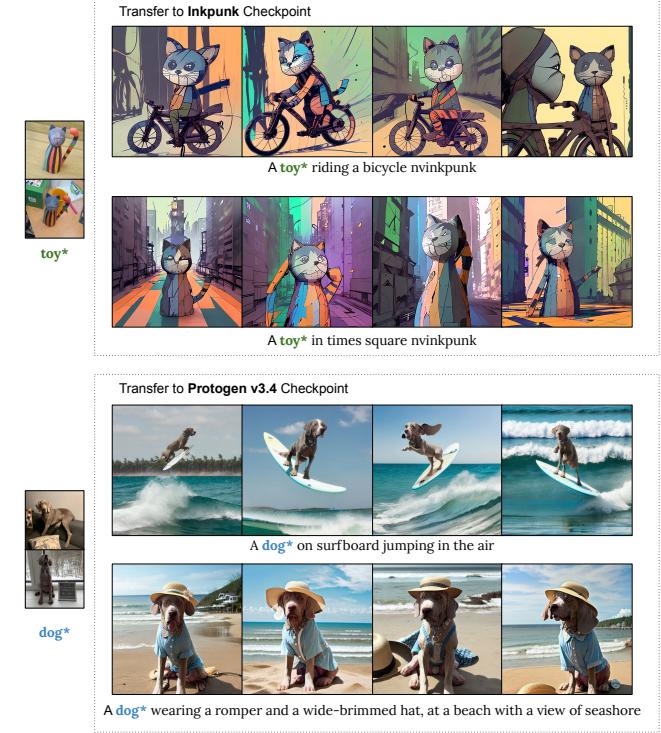


Fig. 12. **Zero-shot transfer to fine-tuned models:** A Perfusion concept trained using a vanilla diffusion-model can generalize to fine-tuned variants.



Fig. 13. **Locking to unusual super-categories:** A concept “inherits” the qualitative outline of the supercategory it is locked onto. For example, the pot as a shoe or a clock, and cat becomes illuminated when locked to lamp.

while hurting the textual alignment. Interestingly, Local Key-Locked training reveals an “over-generalization” phenomenon. As training progresses the concept learns supercategory (latent) features, improving textual alignment but sacrificing visual fidelity. The right panel of Figure 11 displays qualitative examples, where longer training makes the concept “in outer space” increasingly toy-like. It is worth noting that reconstruction prompts have better visual fidelity as the concept is trained using such prompts. This suggests that the learned supercategory features are latent. These findings were expected, as the learned V features propagate to the Q features in

the next denoising step and the Q features should align with the locked supercategory K features due to their inner product when calculating the attention map. Hence, we expected that the V features should learn to encode latent properties of the supercategory in order to improve Q-K alignment. Finally, the example of the cat-toy demonstrates a failure case where the V features pick up the supercategory features too early during training, before the concept has completed learning its fine-grained characteristics.

## 7 CONCLUSIONS AND LIMITATIONS

We have presented *Perfusion*, a novel T2I personalization method that combines high visual fidelity with improved textual alignment. Our approach, which utilizes a gated rank-1 method, provides control over the influence of learned concepts during inference time, enables combination of multiple concepts, and results in a small model size. Importantly, the key-locking technique leads to novel qualitative results compared to traditional approaches.

**Limitations and future work.** We find that the choice of supercategory word to lock onto may sometime produce “over-generalization” effects when using the concept in a new prompt. For instance as shown in Figure 11, setting the concept of a toy-cat as a “toy” may encourage learning to generate it in a childish style in new prompts, while scarifying its visual fidelity. Additionally, Figure 13 demonstrates that locking concepts to atypical super-categories results in the concepts adopting some characteristics of that atypical super-category. A second limitation is that combining concepts requires a great deal of prompt engineering. Interestingly, we found it was easier to succeed on prompts that were suggested by CD, than with prompts that we devised. We believe that there is much of a headroom for improvement in this task.

## ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers of SIGGRAPH 2023 for their valuable feedback, which greatly improved the final version. Additionally, we would like to thank Assaf Hallak, Xun Huang, Jim Fan, and Eli Meirom for providing insightful feedback on an earlier draft of this manuscript.

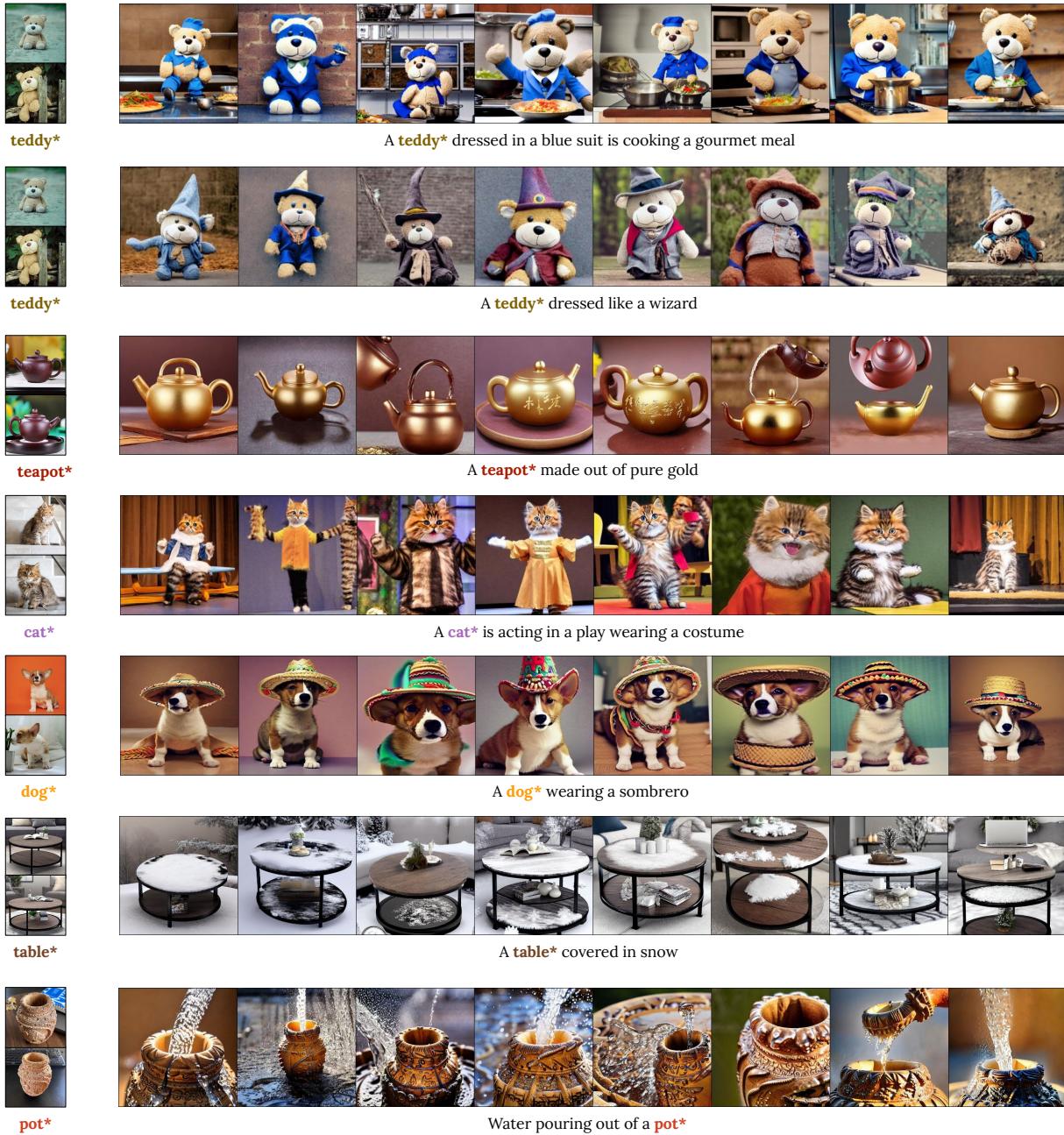


Fig. 14. **Uncurated samples of image variations with text guided prompts.** We observe that a batch size of 8 is typically sufficient to ensure several good samples.

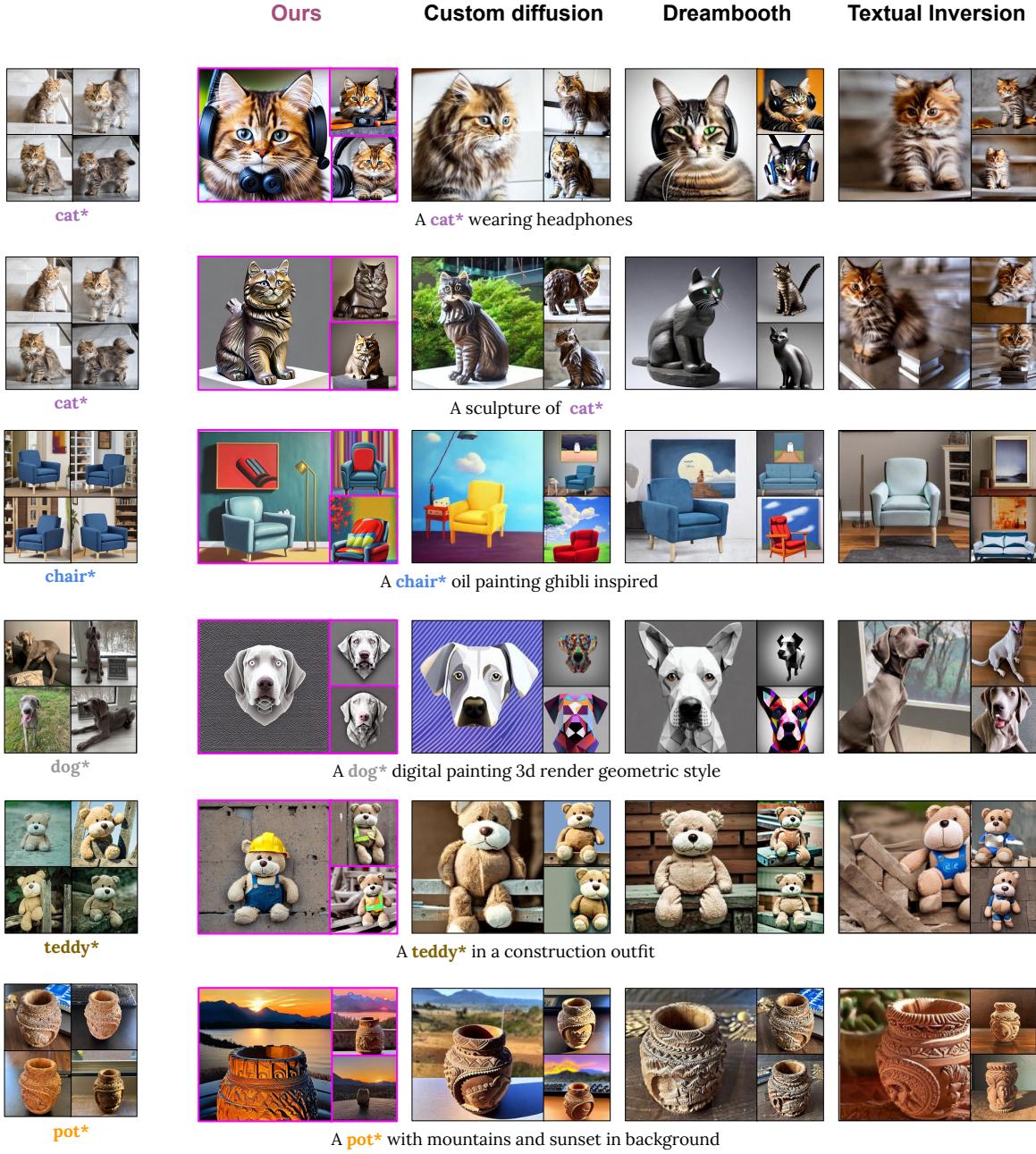


Fig. 15. **Supplement generation results with single concept examples.** For each concept, we show exemplars from our training set, along with generated images, their conditioning texts and comparisons to Custom-Diffusion (CD), Dreambooth (DB), and Textual-Inversion (TI) baselines. All prompts are taken from the CD paper, as well as the baseline generations.

## REFERENCES

- Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4432–4441.
- Harsh Agrawal, Eli A. Meirom, Yuval Atzmon, Shie Mannor, and Gal Chechik. 2021. Known unknowns: Learning novel concepts using reasoning-by-elimination. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 161)*, Cassio de Campos and Marloes H. Maathuis (Eds.). PMLR, 504–514.
- Yuval Atzmon, Uri Shalit, and Gal Chechik. 2015. Learning sparse metrics, one feature at a time. In *Feature Extraction: Modern Questions and Challenges*. PMLR, 30–48.
- Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18208–18218.
- 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. 2022. eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324* (2022).
- David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. Paint by Word. *arXiv:2103.10951 [cs.CV]*
- David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. 2020. Rewriting a Deep Generative Model. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2022. InstructPix2Pix: Learning to Follow Image Editing Instructions. *ArXiv abs/2211.09800* (2022).
- Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. 2022. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations. In *European Conference on Computer Vision (ECCV)*.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*.
- Giannis Daras and Alexandros G. Dimakis. 2022. Multiresolution Textual Inversion. *ArXiv abs/2211.17115* (2022).
- darkstorm2150. 2022. *Protogen-v3.4*. [https://huggingface.co/darkstorm2150/Protogen\\_v3.4\\_Official\\_Release/tree/main](https://huggingface.co/darkstorm2150/Protogen_v3.4_Official_Release/tree/main)
- Envii. 2022. *InkPunk-v2*. <https://huggingface.co/Envvi/Inkpunk-Diffusion>
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. <https://doi.org/10.48550/ARXIV.2208.01618>
- Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946* (2021).
- Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. *arXiv preprint arXiv:2203.14680* (2022).
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913* (2020).
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. (2022).
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.
- Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-Based Real Image Editing with Diffusion Models. *arXiv preprint arXiv:2210.09276* (2022).
- Nupur Kumar, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2022. Multi-Concept Customization of Text-to-Image Diffusion. *arXiv* (2022).
- Benjamin Leflaiveux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, and Daniel Haziza. 2022. xFormers: A modular and hackable Transformer modelling library. <https://github.com/facebookresearch/xformers>.
- Timo Lüddecke and Alexander S Ecker. 2021. Prompt-Based Multi-Modal Image Segmentation. *arXiv preprint arXiv:2112.10003* (2021).
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems* 36 (2022).
- Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models. *arXiv preprint arXiv:2211.09794* (2022).
- Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. *arXiv preprint arXiv:2103.17249* (2021).
- Suraj Patil and Pedro Cuenca. 2022. HuggingFace DreamBooth Implementation. <https://huggingface.co/docs/diffusers/training/dreambooth>.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752 [cs.CV]*
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. (2022).
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamvar Seyed Ghaseimpour, Burcu Karagol Ayan, Sara Mahdavi, Rapha Goncalo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).
- Dvir Samuel, Yuval Atzmon, and Gal Chechik. 2020. From generalized zero-shot learning to long-tail with class descriptors. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020), 286–295.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).
- Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. 2022. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477* (2022).
- Qicheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung M. Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. 2022. Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models. *ArXiv abs/2212.08698* (2022).
- Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2022. Paint by Example: Exemplar-based Image Editing with Diffusion Models. *ArXiv abs/2211.13227* (2022).
- Zhixing Zhang, Ligong Han, Arna Ghosh, Dimitris N. Metaxas, and Jian Ren. 2022. SINE: SINGLE Image Editing with Text-to-Image Diffusion Models. *ArXiv abs/2212.04489* (2022).

## Appendix: Key-Locked Rank One Editing for Text-to-Image Personalization

### A LEMMA: DERIVATION OF THE EXPRESSION FOR THE WEIGHT UPDATE OF ROME

In this section we provide a Lemma that reorganizes and calculates the output of a layer affected by equation Eq. (1) repeated below for convenience.

$$\hat{W} = W + \Lambda(C^{-1}\mathbf{i}_*)^T,$$

where  $\Lambda = (\mathbf{o}_* - W\mathbf{i}_*)/[(\mathbf{i}_*^T(C^{-1})^T\mathbf{i}_*)]$ ,  $C$  is a constant positive definite matrix, that is pre-cached.

**Lemma:** Given an input  $\mathbf{e}_m$ , the output  $h$  of an edited layer is:

$$h = W\mathbf{e}_m^\perp + \mathbf{o}_* \text{sim}(\mathbf{i}_*, \mathbf{e}_m) / \|\mathbf{i}_*\|_{C^{-1}}^2, \quad (5)$$

where (1)  $\text{sim}(\mathbf{i}_*, \mathbf{e}_m) := \mathbf{i}_*^T(C^{-1})^T\mathbf{e}_m$  measures the similarity of  $\mathbf{e}_m$  with  $\mathbf{i}_*$  in a metric space defined by  $C^{-1}$ , (2)  $\|\mathbf{i}_*\|_{C^{-1}}^2 := \text{sim}(\mathbf{i}_*, \mathbf{i}_*)$  measures the energy of  $\mathbf{i}_*$  in the same metric space. (3)  $\mathbf{e}_m^\perp := (\mathbf{e}_m - \mathbf{i}_* \text{sim}(\mathbf{i}_*, \mathbf{e}_m) / \|\mathbf{i}_*\|_{C^{-1}}^2)$  is the component of  $\mathbf{e}_m$  that is orthogonal to  $\mathbf{i}_*$  in the metric space.

**Proof:**

We now proceed to prove the Lemma.

Let  $h$  be the output of the edited layer given an input  $\mathbf{e}_m$ . From equation Eq. (1), we have

$$h = \hat{W}\mathbf{e}_m = W\mathbf{e}_m + \Lambda(C^{-1}\mathbf{i}_*)^T\mathbf{e}_m.$$

Notice that the term  $(C^{-1}\mathbf{i}_*)^T\mathbf{e}_m$  is the similarity of  $\mathbf{e}_m$  with  $\mathbf{i}_*$  in the metric space defined by  $C^{-1}$ , denoted by  $\text{sim}(\mathbf{i}_*, \mathbf{e}_m) = \mathbf{i}_*^T(C^{-1})^T\mathbf{e}_m$ . Thus, we can rewrite the above equation as

$$h = W\mathbf{e}_m + \Lambda \text{sim}(\mathbf{i}_*, \mathbf{e}_m).$$

To simplify the above expression, we examine the term  $\Lambda$ . Recall that  $\Lambda = (\mathbf{o}_* - W\mathbf{i}_*)/[(\mathbf{i}_*^T(C^{-1})^T\mathbf{i}_*)]$ . Thus, we can substitute and rearrange the above equation to obtain

$$\begin{aligned} h &= W\mathbf{e}_m + \frac{\mathbf{o}_* - W\mathbf{i}_*}{\mathbf{i}_*^T(C^{-1})^T\mathbf{i}_*} \text{sim}(\mathbf{i}_*, \mathbf{e}_m) \\ &= W\mathbf{e}_m + \frac{\mathbf{o}_*}{\mathbf{i}_*^T(C^{-1})^T\mathbf{i}_*} \text{sim}(\mathbf{i}_*, \mathbf{e}_m) - \frac{W\mathbf{i}_*}{\mathbf{i}_*^T(C^{-1})^T\mathbf{i}_*} \text{sim}(\mathbf{i}_*, \mathbf{e}_m). \end{aligned}$$

We now define  $\|\mathbf{i}_*\|_{C^{-1}}^2 = \mathbf{i}_*^T(C^{-1})^T\mathbf{i}_*$  for brevity. With this, we can rewrite the above equation as

$$h = W\mathbf{e}_m + \frac{\mathbf{o}_*}{\|\mathbf{i}_*\|_{C^{-1}}^2} \text{sim}(\mathbf{i}_*, \mathbf{e}_m) - \frac{W}{\|\mathbf{i}_*\|_{C^{-1}}^2} \mathbf{i}_* \text{sim}(\mathbf{i}_*, \mathbf{e}_m).$$

To further simplify the above expression, we define  $\mathbf{e}_m^\perp = (\mathbf{e}_m - \mathbf{i}_* \text{sim}(\mathbf{i}_*, \mathbf{e}_m) / \|\mathbf{i}_*\|_{C^{-1}}^2)$ . This is the component of  $\mathbf{e}_m$  that is orthogonal to  $\mathbf{i}_*$  in the metric space defined by  $C^{-1}$ . With this, we can rewrite the above equation as

$$h = W\mathbf{e}_m^\perp + \frac{\mathbf{o}_*}{\|\mathbf{i}_*\|_{C^{-1}}^2} \text{sim}(\mathbf{i}_*, \mathbf{e}_m),$$

as stated in the Lemma.

Namely, the left term nulls the “energy” of  $\mathbf{i}_*$  from  $\mathbf{e}_m$  and passes the rest ( $\mathbf{e}_m^\perp$ ) through the original matrix  $W$ . Then, the right term grossly assigns that energy in the direction of the  $\mathbf{o}_*$  component.

### B DERIVATION OF $\mathbf{e}_m^{\perp J}$ , THE ORTHOGONAL COMPONENT FOR MULTIPLE CONCEPTS

To calculate the orthogonal component of  $\mathbf{e}_m$  for multiple concepts, we start by projecting  $\mathbf{e}_m$  and the set of target-inputs  $\{\mathbf{i}_*^j\}_{j=1..J}$  to the metric space of  $C^{-1}$ , using a Cholesky decomposition  $C^{-1} = LL^T$

$$\begin{aligned} \widetilde{\mathbf{e}_m} &= L^T \mathbf{e}_m \\ \widetilde{\mathbf{i}_*^j} &= L^T \mathbf{i}_*^j, \end{aligned} \quad (6)$$

where we denote by “~” every vector in the metric space.

Next, we use a QR decomposition to find an orthonormal spanning basis for  $\{\widetilde{\mathbf{i}_*^j}\}_{j=1..J}$ . We denote this basis by  $\{\widetilde{\mathbf{u}}_j\}_{j=1..J}$ .

Then, the orthogonal component of  $\widetilde{\mathbf{e}_m}$  in this metric space is

$$\widetilde{\mathbf{e}_m^{\perp J}} = \widetilde{\mathbf{e}_m} - \sum_{j \in 1..J} \widetilde{\mathbf{u}}_j \langle \widetilde{\mathbf{u}}_j, \widetilde{\mathbf{e}_m} \rangle, \quad (7)$$

where  $\langle \widetilde{\mathbf{u}}_j, \widetilde{\mathbf{e}_m} \rangle$  is simply their dot product  $\widetilde{\mathbf{u}}_j^T \cdot \widetilde{\mathbf{e}_m}$ .

Projecting the last expression back to text encoder space using the inverse of the Cholesky root  $L^{T-1}$ , multiplied from the left yields

$$\mathbf{e}_m^{\perp J} = \mathbf{e}_m - \sum_{j \in 1..J} u_j \langle \widetilde{\mathbf{u}}_j, \widetilde{\mathbf{e}_m} \rangle, \quad (8)$$

where  $u_j = L^{T-1} \widetilde{\mathbf{u}}_j$ .

Finally, note that  $\langle \widetilde{\mathbf{u}}_j, \widetilde{\mathbf{e}_m} \rangle = \text{sim}(u_j, \mathbf{e}_m)$ , because

$$\langle \widetilde{\mathbf{u}}_j, \widetilde{\mathbf{e}_m} \rangle = \widetilde{\mathbf{u}}_j^T \cdot \widetilde{\mathbf{e}_m} = u_j L L^T \mathbf{e}_m = \text{sim}(u_j, \mathbf{e}_m) \quad (9)$$

Which brings us to our final expression for  $\mathbf{e}_m^{\perp J}$

$$\mathbf{e}_m^{\perp J} = \mathbf{e}_m - \sum_{j \in 1..J} u_j \text{sim}(u_j, \mathbf{e}_m) \quad (10)$$

#### B.1 The special case of a single concept

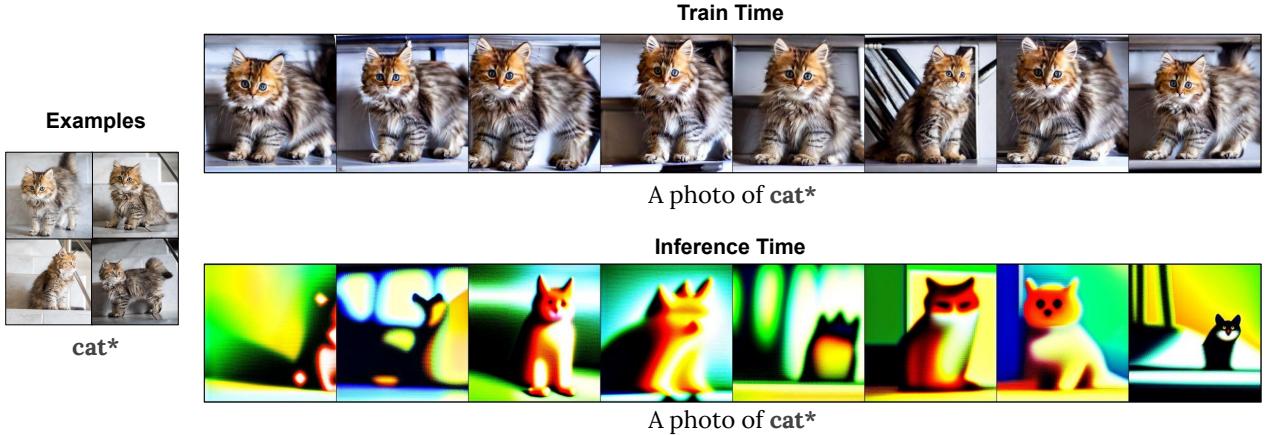
As a final step, we prove that when there is only a single concept, denoted by  $J = 1$ , then  $\mathbf{e}_m^{\perp J} = (\mathbf{e}_m - \mathbf{i}_* \text{sim}(\mathbf{i}_*, \mathbf{e}_m) / \|\mathbf{i}_*\|_{C^{-1}}^2)$ .

By definition,  $\widetilde{\mathbf{u}}_1 = L^T \mathbf{i}_* / \|L^T \mathbf{i}_*\|$ , and thus  $u_1 = \mathbf{i}_* / \|L^T \mathbf{i}_*\|$ . Substituting this into Equation Eq. (10), we obtain:

$$\mathbf{e}_m^{\perp J} = \mathbf{e}_m - \frac{\mathbf{i}_*}{\|L^T \mathbf{i}_*\|} \text{sim}\left(\frac{\mathbf{i}_*}{\|L^T \mathbf{i}_*\|}, \mathbf{e}_m\right) = \mathbf{e}_m - \frac{\mathbf{i}_* \text{sim}(\mathbf{i}_*, \mathbf{e}_m)}{\|L^T \mathbf{i}_*\|^2} \quad (11)$$

To complete the proof, we note that  $\|L^T \mathbf{i}_*\|^2 = \mathbf{i}_*^T C^{-1} \mathbf{i}_* = \|\mathbf{i}_*\|_{C^{-1}}^2$ , by definition.

This concludes the proof that  $\mathbf{e}_m^{\perp J} = (\mathbf{e}_m - \mathbf{i}_* \text{sim}(\mathbf{i}_*, \mathbf{e}_m) / \|\mathbf{i}_*\|_{C^{-1}}^2)$  when there is only a single concept,  $J = 1$ .



**Fig. A.1. Train-Inference Mismatch:** As presented in Challenge 1, when training the model with ROME’s naive solution, we observe corrupted generations during inference time. We show that the low quality of generated images is the result of the mismatch between the objective during inference and training time.

## C ADDITIONAL TECHNICAL DETAILS

**Estimating  $C$ :**  $C$  is constant matrix that is pre-cached by estimating the uncentered covariance of  $i$  from a sample of 100K LAION image captions,  $C^{-1}$  is its inverse, and both are positive definite (and symmetric).

**Global Key Locking:** Here we describe our inference time approach called *global key-locking*.

- (1) Given a key-locked trained concept and a prompt that includes it, we start by making a forward pass through the text encoder and calculate the K and V sequence activations that this prompt elicits, denoted by  $K$ , and  $V$ .
- (2) We replace the concept word in the prompt by its superclass word and make another forward pass through the text encoder. This time, we only calculate the K sequence activations that the super-class prompt elicits, denoted by  $\tilde{K}$ .
- (3) We override  $K$  with  $\tilde{K}$  and proceed to sample the image as usual.

We refer to this variant as *global key-locking*, because it locks the K pathway of the entire prompt.

## D PSEUDO CODE

Algorithm 1 shows a pseudo code for the Rank-One Edit Neural Module. This module implements a single-layer edit and the same implementation is used for all layers. It is presented in a PyTorch module style for convenience. The module replaces the keys (K) and values (V) projection modules in the cross-attention layers during training time. It implements Eq. (3) jointly with an online estimation of  $i_*$ . The only optimized variable is  $o_*$ , when  $o_*.require\_grad$  is *True*. For inference with a single concept, we can simply disable the online estimation of  $i_*$ .

Where “@” operator denotes matrix multiplication,  $init\_input$  is the encoding of a prompt saying “A photo of a <superclass\_word>”,  $input$  is the encoding  $e$  of an entire prompt,  $W$  denotes the pre-trained layer weights,  $concept\_token\_idx$  is the index of the concept

---

**ALGORITHM 1:** Rank-One Edit Module for a Single Concept (Eq. 3)

---

```

Init Constants:  $W, C^{-1}, \beta, \tau,$ 
def init(init_input, is_key_locked, concept_token_idx):
     $i_* = init\_input[concept\_token\_idx]$ 
     $o_* = W @ init\_input;$ 
    # Only train  $o_*$  of  $V$  pathway
     $o_*.require\_grad = is\_key\_locked$ 

def forward(input, concept_token_idx):
    # online estimation of  $i_*$ :
    concept_encoding = input[concept_token_idx]
     $i_* = 0.99 * i_* + 0.01 * concept\_encoding$ 

    # Eq. 3 components:
     $i_*\_energy = (C^{-1} @ i_*)^T @ i_*$ 
    sim = (input@( $C^{-1} @ i_*$ ) $^T$ )
    sigmoid_term =  $\sigma\left(\frac{(sim/i_*\_energy)-\beta}{\tau}\right)$ 
    # calculating  $W e_m^\perp$ :
    W_em_orthogonal_term =
         $W @ input - (sim \cdot (W @ i_*) / i_*\_energy)$ 
    # Eq. 3:
     $h = W\_em\_orthogonal\_term + sigmoid\_term @ o_*$ 
    return h
  
```

---

word  $S^*$ ,  $C^{-1}$  is the inverse of the constant matrix  $C$ , and  $\beta, \tau$  are the sigmoid hyper-parameters. Keep in mind that because Eq. (3) is valid for all  $m$ , this module processes the entire prompt simultaneously.

## E TRAINING CLASSES

- cat toy (from Textual-inversion)
- headless sculpture (from Textual-inversion)
- cat (from Custom-Diffusion)
- chair (from Custom-Diffusion)

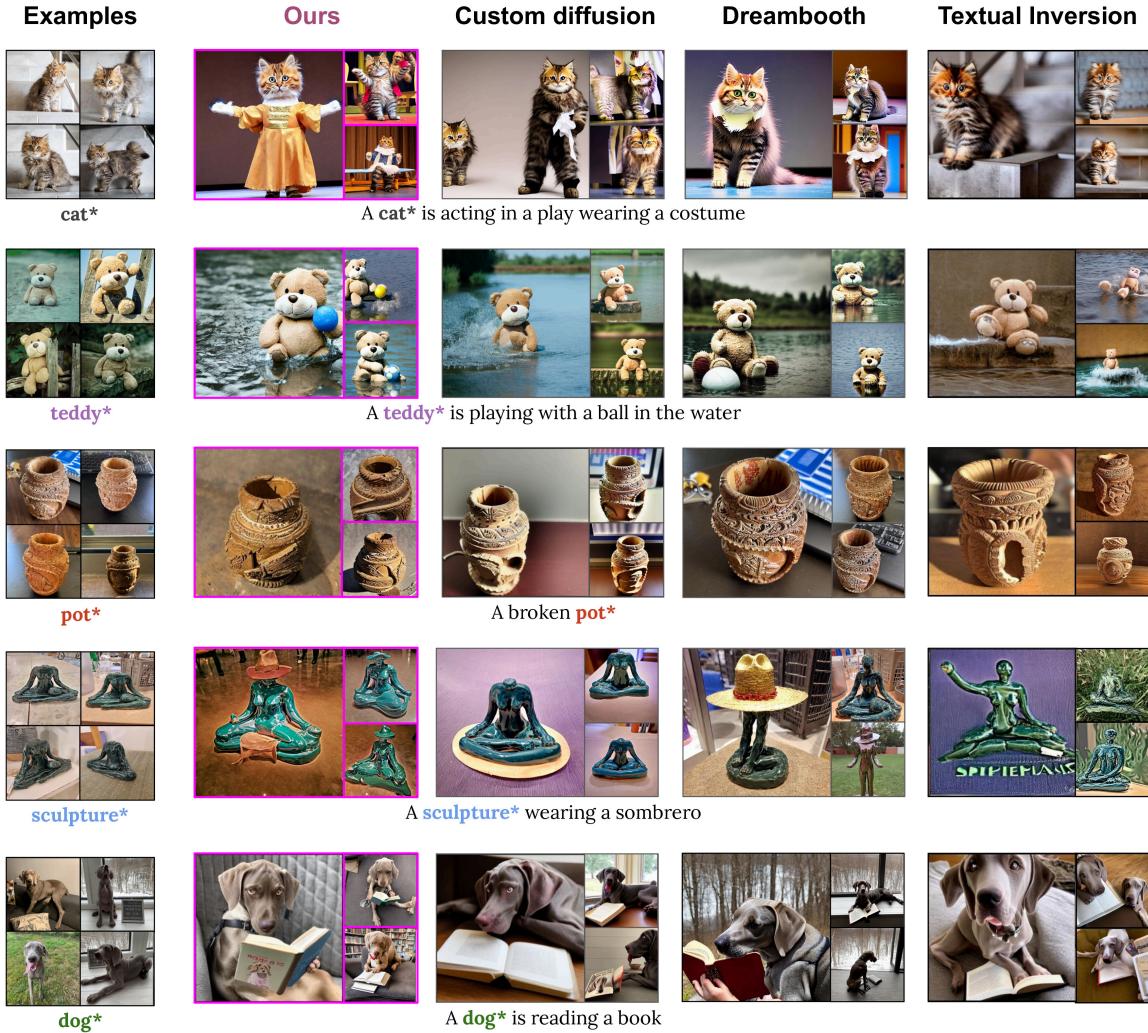


Fig. A.2. Qualitative comparison including Textual Inversion.

- wooden pot (from Custom-Diffusion)
- dog (from Custom-Diffusion)
- teadybear (from Custom-Diffusion)
- tortoise plushy (from Custom-Diffusion)
- puppy (from DreamBooth)
- sunglasses (from DreamBooth)
- teapot (similar to DreamBooth teapot)

## F SUCCESS AND FAILURES OF GLOBAL LOCKING

Image-similarity of global locking may appear worse than local locking, but it is not always the case. We find that global locking struggles with uniquely shaped objects, but is successful with common objects like pets, teddy bears, and chairs. Figure A.4 shows

how global locking struggles with uniquely shaped objects, but is successful with everyday concepts. In cases where global locking is successful, we find that it allows to generate rich scenes and portrays better the nuances of the object attributes or activities.

## G TRAINING PROMPTS

We used the following set of prompts for training:

- a photo of a  $S^*$
- a good photo of a  $S^*$
- the photo of a  $S^*$
- a good photo of the  $S^*$
- image of a  $S^*$
- image of the  $S^*$

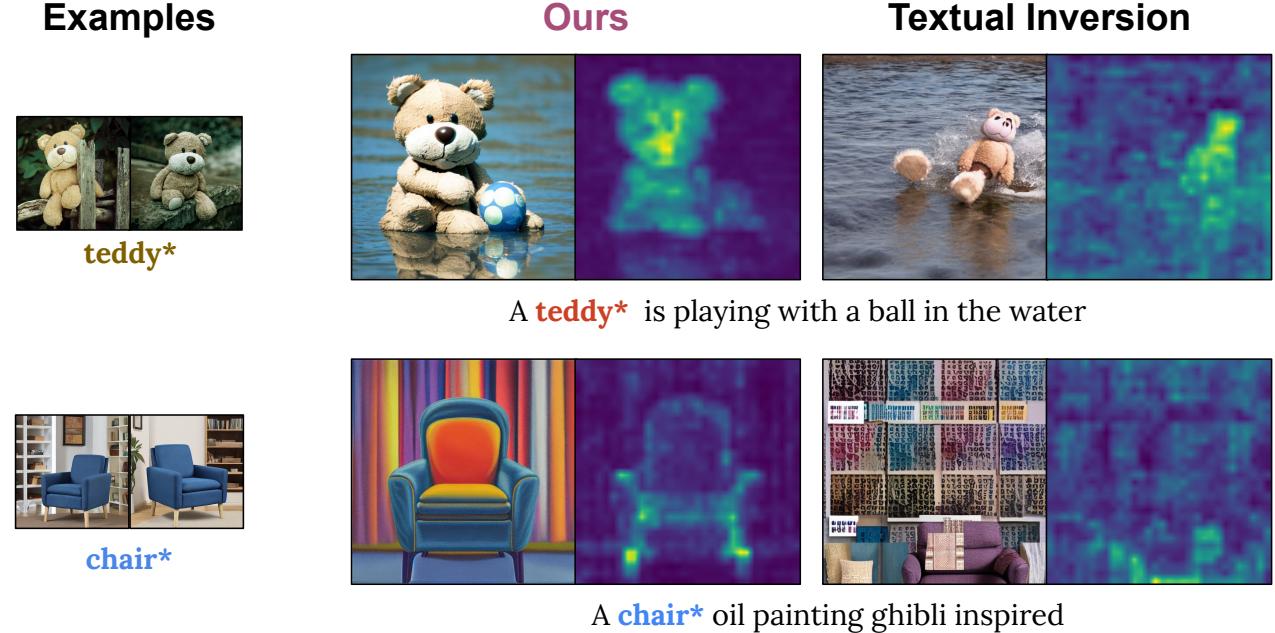


Fig. A.3. Additional results of attention overfit in Textual inversion

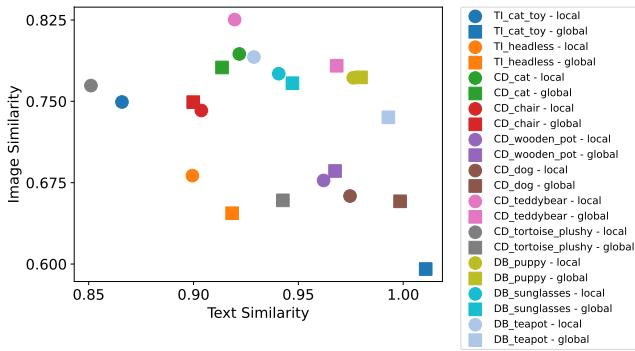


Fig. A.4. Success and failures of global locking: Circles - using local locking. Square - using global locking. Global locking struggles with classes of uniquely shaped object like CD\_tortoise\_plushy. However it is successful with classes of everyday concepts like CD\_dog.

- A photograph of  $S^*$
- A  $S^*$  shown in a photo,
- A photo of  $S^*$

## H QUALITATIVE FIGURES PARAMETERS CHOICE

We describe the inference-time parameters we chose for each prompt in qualitative figures:

Figure 1:

- "A Teapot\* made out of {pure gold, yarn, glass}" = Lock: Global, Train Bias: 0.75, Infer Bias: 0.45, Infer temp: 0.1
- "A Teapot\* oil painting ghibli inspired" = Lock: Global, Train Bias: 0.75, Infer Bias: 0.45, Infer temp: 0.1
- "A Teddy\* dressed as a {samurai, superhero, wizard}" = Lock: Global, Train Bias: 0.75, Infer Bias: 0.45, Infer temp: 0.1
- "A Teddy\* dressed in a blue suit is cooking a gourmet meal" = Lock: local, Train Bias: 0.75, Infer Bias: 0.75, Infer temp: 0.1
- "A Teddy\* is sailing inside a Teapot\* in a lake" = Lock: local, Train Bias: 0.75, Infer Bias: 0.55, Infer temp: 0.15
- "A Teddy\* sitting by the fire with a Teapot\* on a chilly night" = Lock: local, Train Bias: 0.75, Infer Bias: 0.55, Infer temp: 0.15
- "Painting of a Teddy\* is sitting next to a Teapot\* on a picnic" = Lock: local, Train Bias: 0.75, Infer Bias: 0.55, Infer temp: 0.15
- "The Teddy\* in engraved on a Teapot\*" = Lock: local, Train Bias: 0.75, Infer Bias: 0.55, Infer temp: 0.15

**Figure 4:**

- "A Sculpture\* wearing a sombrero" = Lock: local, Train Bias: 0.55, Infer Bias: 0.65, Infer temp: 0.15
- "A Cat\* is acting in a play wearing a costume" = Lock: global, Train Bias: 0.75, Infer Bias: 0.65, Infer temp: 0.15
- "A Teddy\* is playing with a ball in the water" = Lock: local, Train Bias: 0.75, Infer Bias: 0.75, Infer temp: 0.1
- "A broken Pot\*\*" = Lock: global, Train Bias: 0.75, Infer Bias: 0.45, Infer temp: 0.15
- "A Dog\* is reading a book" = Lock: global, Train Bias: 0.75, Infer Bias: 0.45, Infer temp: 0.15

**Figure 5:**

- "Watercolor painting of Cat\* sitting on Chair\*" = Lock: local, Train Bias: 0.6, Infer Bias: 0.6, Infer temp: 0.1
- "Photo of a Table\* and the Chair\*" = Lock: local, Train Bias: 0.7, Infer Bias: 0.75, Infer temp: 0.1
- "The Cat\* playing with a Pot\* in a garden" = Lock: local, Train Bias: 0.75, Infer Bias: 0.4, Infer temp: 0.1
- "A Teddy\* is sitting in the garden and wearing the Sunglasses\*" = Lock: local, Train Bias: 0.75, Infer Bias: 0.55, Infer temp: 0.15

## I USER STUDY DETAILS

We evaluate the models by two user studies conducted with Amazon Mechanical Turk. In the first, raters were asked to rank a set of images generated by the three methods (*Perfusion*, CD, DB). In the second, we compared *Perfusion* to “vanilla” stable-diffusion (SD) in order to study whether *Perfusion* harms the generative prior.

### I.1 Personalization Method Comparison

For the first study, in each trial, raters were shown two train set images of a concept and a prompt description. They were then asked to examine a set of three generated images, one from each method, and assign each image a unique rank from 1 to 3 (best), based on how well they portrayed the concept according to the description. Method order was randomized in each trial. Figure A.5 illustrates the experimental framework that was used in trials. Figure A.6 displays the examples provided to help guide raters through the instructions.

We had 2104 trials from 11 concepts, with an average of 24 prompts per concept, and 8 trials per prompt. We used all the prompts from the challenging subset of *per-group* prompts. For *Perfusion* method images, we select the runtime variant with the highest harmonic mean for text and visual similarities in each *prompt*. Nonetheless, both methods require the same computational resources. Reported rank scores are balanced “per-class” [Samuel et al. 2020]. Namely, we compute the mean score per concept class, and then uniformly average between all class scores.

We paid \$0.12 per trial. To maintain the quality of the queries, we only picked raters with AMT “masters” qualification, demonstrating a high degree of approval rate over a wide range of tasks. Furthermore, we also conducted a qualification test on the prescreened pool of raters, consisting of a few curated trials that were very simple. In each qualification trial, we included one well generated image that depict the concept as described by the prompt, one image generated from vanilla stable-diffusion, and one image from the concept training examples. We only qualified raters who had completed a

minimum of 5 trials out of a pool of 11 trials with perfect scores. The reason for not having all raters complete all trials is that AMT assigns them randomly. In one of our qualification trials, which is demonstrated in Figure A.5, a significant proportion of raters in the prescreened pool did not perform well. As a result, we replicated the task and made sure that all qualified raters had successfully completed it.

### I.2 Generative Prior Preservation

For the second study, in each trial, raters were given a prompt description and two generated images, one by *Perfusion* and another by SD. They were asked to rate how realistic each image looks, based on the text description. The participants were instructed to assign each image a score between 1 to 3 (best). Unlike the first study, the raters were allowed to give the same score for both images. Method order was randomized in each trial. We had the same number of trials, prompts and concepts as in the first study. Additionally, we utilized the same images generated for *Perfusion*. Figure A.7 illustrates the experimental framework that was used in trials. Figure A.8 displays the examples provided to help guide raters through the instructions.

We paid \$0.1 per trial. To maintain the quality of the queries, we only picked raters with AMT “masters” qualification. Furthermore, we also executed a qualification test with a few curated trials that are very simple. We only qualified raters who had completed all trials with perfect scores and had completed at least 5 trials.

Rank how well each image portray "my-cat" according to the description.



**Here is "my-cat":**  
**Description:** Painting of two my-cat on a boat

#### Instructions

- **Read the instructions carefully**
- You are given two examples of an object, and a description to portray that object in a picture.
- Your mission is to examine a set of images and rank how well they follow the description.
- **IMPORTANT!** : You should assign a **different rank** (1..3) for each image.
- Make sure to **pay attention to the description**. For example, in "my teddy is drinking from a cup", rank higher images where the teddy is **really drinking** from the cup.
- Make sure to **pay attention to the object examples**. Rank higher images that their object **looks similar** to the examples.
- **Do not** select the **same score** for two images. Your HIT may be rejected for such a selection.
- **Do not** select a **score of 0**. Your HIT may be rejected for such a selection.
- **Do not** abuse the system, we take measures to spot that.
- Please **take a look at the examples** with their solution (on the top tabs of this page).



1 = worst, ..., 3 = best



1 = worst, ..., 3 = best



1 = worst, ..., 3 = best

Fig. A.5. One qualification trial of the method comparison user study. This was the hardest for the prescreened pool of raters, and we made sure all qualified users were tested with it.

Task    Example 1    Example 2

Rank how well each image portray "my-teapot" according to the description.



**Here is "my-teapot":**  
**Description:** A photo of golden my-teapot

In this example, when evaluating "my-teapot" images, it is important to consider both the physical appearance of the teapot and the description provided. The middle image is a golden version of "my-teapot" and should receive the highest ranking (3). The right image does not comply with the description as it is not "my-teapot", despite being golden. The left image also does not comply with the description as it is "my-teapot," but not golden.



Task    Example 1    Example 2

Rank how well each image portray "my-teapot" according to the description.



**Here is "my-dog":**  
**Description:** A painting of my-dog in the style of Monet

In this example, when evaluating "my-dog" images, it is important to consider both the physical appearance of the dog and the description provided. The middle image is a painting in the style of Monet of "my-dog" and should receive the highest ranking (3). The right image does not comply with the description as it is not "my-dog", despite being a painting in the style of Monet. The left image also does not comply with the description as it is "my-dog," but not a painting in the style of Monet.



Fig. A.6. Examples provided in the method comparison user study.

Rate how realistic each image looks, based on the text description  
**Description:** A dog wearing a sombrero

### Instructions

- Read the instructions carefully
- You are given a textual description.
- Your mission is to examine a set of images and rate how realistic each image looks, based the description.
- **IMPORTANT!!** : You should assign each image a score between 1 (Worst) and 3 (Best).
- If the 2 images appears to be at the same level of realism, you can give them the same score.
- **Do not** select a **score of 0**. Your HIT may be rejected for such a selection.
- **Do not** abuse the system, we take measures to spot that.



1 = worst, ..., 3 = best



1 = worst, ..., 3 = best

Fig. A.7. One trial of the prior preservation user study.

Task    Example 1    Example 2

Rank how well each image portray a text description  
**Description:** A teddy playing with a ball in the water

In this example, it is important to consider the realism of the images. The left image seems realistic, like a teddy bear would look when playing with a ball in the water, so it should receive a high score (3). The right image does not seem as realistic, it doesn't look like a real teddy bear that plays with a ball, and it should receive a lower score (1)



Task    Example 1    Example 2

Rank how well each image portray a text description  
**Description:** A cat acting in a play wearing a costume

In this example, it is important to consider the realism of the images. The left image doesn't seem like a real cat, it is blurry and its shape is not realistic, so it should receive a lower score (1 or 2). The right image does look like a real cat acting in a play, so it should receive a higher score (3).



Fig. A.8. Examples provided in the prior preservation user study.