An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Rinon Gal[1,2*] **Yuval Alaluf**[1] **Yuval Atzmon**[2] **Or Patashnik**[1]

Amit H. Bermano[1] **Gal Chechik**[2] **Daniel Cohen-Or**[1]

[1]**Tel-Aviv University** [2]**NVIDIA**

Abstract

Text-to-image models offer unprecedented freedom to guide creation through natural language. Yet, it is unclear how such freedom can be exercised to generate images of specific unique concepts, modify their appearance, or compose them in new roles and novel scenes. In other words, we ask: how can we use language-guided models to turn *our* cat into a painting, or imagine a new product based on *our* favorite toy? Here we present a simple approach that allows such creative freedom. Using only 3-5 images of a user-provided concept, like an object or a style, we learn to represent it through new "words" in the embedding space of a frozen text-to-image model. These "words" can be composed into natural language sentences, guiding *personalized* creation in an intuitive way. Notably, we find evidence that a *single* word embedding is sufficient for capturing unique and varied concepts. We compare our approach to a wide range of baselines, and demonstrate that it can more faithfully portray the concepts across a range of applications and tasks.

Our code, data and new words will be available at: `https://textual-inversion.github.io`

Figure 1: (left) We find new pseudo-words in the embedding space of a pre-trained text-to-image model which describe specific concepts. (right) These pseudo-words can be composed into new sentences, placing our targets in new scenes, changing their style or composition, or ingraining them into new products.

Preprint. Under review.

[*] Work was done during an internship at NVIDIA

1 Introduction

In a famous scene from the motion picture "Titanic", Rose makes a request of Jack: "...draw me like one of your French girls". Albeit simple, this request contains a wealth of information. It indicates that Jack should produce a drawing; It suggests that its style and composition should match those of a subset of Jack's prior work; Finally, through a single word, "me", Rose indicates that this drawing should portray a specific, unique subject: Rose herself. In making her request, Rose relies on Jack's ability to reason over these concepts — both broad and specific — and bring them to life in a new creation.

Recently, large-scale text-to-image models (Rombach et al., 2021; Ramesh et al., 2021, 2022; Nichol et al., 2021; Yu et al., 2022; Saharia et al., 2022) have demonstrated an unprecedented capability to reason over natural language descriptions. They allow users to synthesize novel scenes with unseen compositions and produce vivid pictures in a myriad of styles. These tools have been used for artistic creation, as sources of inspiration, and even to design new, physical products (Yacoubian, 2022). Their use, however, is constrained by the user's ability to describe the desired target through text. Turning back to Rose, one could then ask: How might she frame her request if she were to approach one of these models? How could we, as users, ask text-to-image models to craft a novel scene containing a cherished childhood toy? Or to pull our child's drawing from its place on the fridge, and turn it into an artistic showpiece?

Introducing new concepts into large scale models is often difficult. Re-training a model with an expanded dataset for each new concept is prohibitively expensive, and fine-tuning on few examples typically leads to catastrophic forgetting (Ding et al., 2022; Li et al., 2022). More measured approaches freeze the model and train transformation modules to adapt its output when faced with new concepts (Zhou et al., 2021; Gao et al., 2021; Skantze & Willemsen, 2022). However, these approaches are still prone to forgetting prior knowledge, or face difficulties in accessing it concurrently with newly learned concepts (Kumar et al., 2022; Cohen et al., 2022).

We propose to overcome these challenges by *finding* new words in the textual embedding space of pre-trained text-to-image models. We consider the first stage of the text encoding process (Figure 2). Here, an input string is first converted to a set of tokens. Each token is then replaced with its own embedding vector, and these vectors are fed through the downstream model. Our goal is to find new embedding vectors that represent new, specific concepts.

We represent a new embedding vector with a new *pseudo-word* (Rathvon, 2004) which we denote by S_*. This pseudo-word is then treated like any other word, and can be used to compose novel textual queries for the generative models. One can therefore ask for "a photograph of S_* on the beach", "an oil painting of a S_* hanging on the wall", or even compose two concepts, such as "a drawing of S_*^1 in the style of S_*^2". Importantly, this process leaves the generative model untouched. In doing so, we retain the rich textual understanding and generalization capabilities that are typically lost when fine-tuning vision and language models on new tasks.

To find these pseudo-words, we frame the task as one of inversion. We are given a fixed, pre-trained text-to-image model and a small (3-5) image set depicting the concept. We aim to find a single word embedding, such that sentences of the form "A photo of S_*" will lead to the reconstruction of images from our small set. This embedding is found through an optimization process, which we refer to as "Textual Inversion".

We further investigate a series of extensions based on tools typically used in Generative Adversarial Network (GAN) inversion. Our analysis reveals that, while some core principles remain, applying the prior art in a naïve way is either unhelpful or actively harmful.

We demonstrate the effectiveness of our approach over a wide range of concepts and prompts, showing that it can inject unique objects into new scenes, transform them across different styles, transfer poses, diminish biases, and even imagine new products.

In summary, our contributions are as follows:

- We introduce the task of personalized text-to-image generation, where we synthesize novel scenes of user-provided concepts guided by natural language instruction.
- We present the idea of "Textual Inversions" in the context of generative models. Here the goal is to find new pseudo-words in the embedding space of a text encoder that can capture both high-level semantics and fine visual details.

- We analyze the embedding space in light of GAN-inspired inversion techniques and demonstrate that it also exhibits a tradeoff between distortion and editability. We show that our approach resides on an appealing point on the tradeoff curve.
- We evaluate our method against images generated using user-provided captions of the concepts and demonstrate that our embeddings provide higher visual fidelity, and also enable more robust editing.

2 Related work

Text-guided synthesis. Text-guided image synthesis has been widely studied in the context of GANs (Goodfellow et al., 2014). Typically, a conditional model is trained to reproduce samples from given paired image-caption datasets (Zhu et al., 2019; Tao et al., 2020), leveraging attention mechanisms (Xu et al., 2018) or cross-modal contrastive approaches (Zhang et al., 2021; Ye et al., 2021). More recently, impressive visual results were achieved by leveraging large scale auto-regressive (Ramesh et al., 2021; Yu et al., 2022) or diffusion models (Ramesh et al., 2022; Saharia et al., 2022; Nichol et al., 2021; Rombach et al., 2021).

Rather than training conditional models, several approaches employ test-time optimization to explore the latent spaces of a pre-trained generator (Crowson et al., 2022; Murdock, 2021; Crowson, 2021). These models typically guide the optimization to minimize a text-to-image similarity score derived from an auxiliary model such as CLIP (Radford et al., 2021).

Moving beyond pure image generation, a large body of work explores the use of text-based interfaces for image editing (Patashnik et al., 2021; Abdal et al., 2021; Avrahami et al., 2022b), generator domain adaptation (Gal et al., 2021; Kim et al., 2022), video manipulation (Tzaban et al., 2022; Bar-Tal et al., 2022), motion synthesis (Tevet et al., 2022; Petrovich et al., 2022), style transfer (Kwon & Ye, 2021; Liu et al., 2022) and even texture synthesis for 3D objects (Michel et al., 2021).

Our approach builds on the open-ended, conditional synthesis models. Rather than training a new model from scratch, we show that we can expand a frozen model's vocabulary and introduce new pseudo-words that describe specific concepts.

GAN inversion. Manipulating images with generative networks often requires one to find a corresponding latent representation of the given image, a process referred to as *inversion* (Zhu et al., 2016; Xia et al., 2021). In the GAN literature, this inversion is done through either an optimization-based technique (Abdal et al., 2019, 2020; Zhu et al., 2020b; Gu et al., 2020) or by using an encoder (Richardson et al., 2020; Zhu et al., 2020a; Pidhorskyi et al., 2020; Tov et al., 2021). Optimization methods directly optimize a latent vector, such that feeding it through the GAN will re-create a target image. Encoders leverage a large image set to train a network that maps images to their latent representations.

In our work, we follow the optimization approach, as it can better adapt to unseen concepts. Encoders face harsher generalization requirements, and would likely need to be trained on web-scale data to offer the same freedom. We further analyze our embedding space in light of the GAN-inversion literature, outlining the core principles that remain and those that do not.

Diffusion-based inversion. In the realm of diffusion models, inversion can be performed naïvely by adding noise to an image and then de-noising it through the network. However, this process tends to change the image content significantly. Choi et al. (2021) improve inversion by conditioning the denoising process on noised low-pass filter data from the target image. (Dhariwal & Nichol, 2021) demonstrate that the DDIM (Song et al., 2020) sampling process can be inverted in a closed-form manner, extracting a latent noise map that will produce a given real image. In DALL-E 2 (Ramesh et al., 2022), they build on this method and demonstrate that it can be used to induce changes in the image, such as cross-image interpolations or semantic editing. The later relies on their use of CLIP-based codes to condition the model, and may not be applicable to other methods.

Whereas the above works invert a given *image* into the model's latent space, we invert a user-provided *concept*. Moreover, we represent this concept as a new pseudo-word in the model's vocabulary, allowing for more general and intuitive editing.

Personalization. Adapting models to a specific individual or object is a long-standing goal in machine learning research. Personalized models are typically found in the realms of recommendation systems (Benhamdi et al., 2017; Amat et al., 2018; Martinez et al., 2009; Cho et al., 2002) or in federated learning (Mansour et al., 2020; Jiang et al., 2019; Fallah et al., 2020; Shamsian et al., 2021).

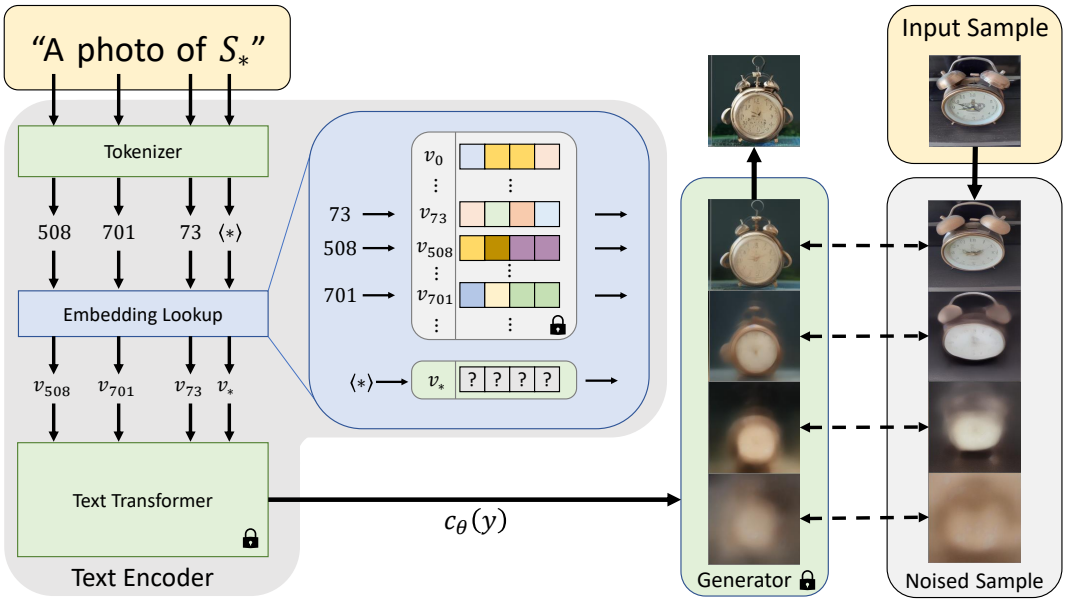


Figure 2: Outline of the text-embedding and inversion process. A string containing our placeholder word is first converted into tokens (*i.e.* word or sub-word indices in a dictionary). These tokens are converted to continous vector representations (the "embeddings", v). Finally, the embedding vectors are transformed into a single conditioning code $c_\theta(y)$ which guides the generative model. We optimize the embedding vector v_* associated with our pseudo-word S_*, using a reconstruction objective.

More recently, personalization efforts can also be found in vision and graphics. There it is typical to apply a delicate tuning of a generative model to better reconstruct specific faces or scenes (Bau et al., 2019; Roich et al., 2021; Alaluf et al., 2021; Dinh et al., 2022; Cao et al., 2022; Nitzan et al., 2022).

Most relevant to our work is PALAVRA (Cohen et al., 2022), which leverages a pre-trained CLIP model for retrieval and segmentation of personalized objects. PALAVRA identifies pseudo-words in the textual embedding space of CLIP that refer to a specific object. These are then used to describe images for retrieval, or in order to segment specific objects in a scene. However, their task and losses are both discriminative, aiming to separate the object from other candidates. As we later show (Figure 5), their approach fails to capture the details required for plausible reconstructions or synthesis in new scenes.

3 Method

Our goal is to enable language-guided generation of new, user-specified concepts. To do so, we aim to encode these concepts into an intermediate representation of a pre-trained text-to-image model. Ideally, this should be done in a manner that would allow us to leverage the rich semantic and visual prior represented by such a model, and use it to guide intuitive visual transformations of the concepts.

It is natural to search for candidates for such a representation in the word-embedding stage of the text encoders typically employed by text-to-image models. There, the discrete input text is first converted into a continuous vector representation that is amenable to direct optimization.

Prior work has shown that this embedding space is expressive enough to capture basic image semantics (Cohen et al., 2022; Tsimpoukelli et al., 2021). However, these approaches leveraged contrastive or language-completion objectives, neither of which require an in-depth visual understanding of the image. As we demonstrate in Section 4, those methods fail to accurately capture the appearance of the concept, and attempting to employ them for synthesis leads to considerable visual corruption. Our goal is to find pseudo-words that can guide *generation*, which is a *visual* task. As such, we propose to find them through a *visual* reconstruction objective.

Below, we outline the core details of applying our approach to a specific class of generative models — Latent Diffusion Models (Rombach et al., 2021). In Section 5, we then analyze a set of extensions to this approach,

motivated by GAN-inversion literature. However, as we later show, these additional complexities fail to improve upon the initial representation, presented here.

Latent Diffusion Models. We implement our method over Latent Diffusion Models (LDMs) (Rombach et al., 2021), a recently introduced class of Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) that operate in the latent space of an autoencoder.

LDMs consist of two core components. First, an autoencoder is pre-trained on a large collection of images. An encoder $\mathcal{E}$ learns to map images $x \in \mathcal{D}_x$ into a spatial latent code $z = \mathcal{E}(x)$, regularized through either a KL-divergence loss or through vector quantization (Van Den Oord et al., 2017; Agustsson et al., 2017). The decoder D learns to map such latents back to images, such that $D\left(\mathcal{E}(x)\right) \approx x$.

The second component, a diffusion model, is trained to produce codes within the learned latent space. This diffusion model can be conditioned on class labels, segmentation masks, or even on the output of a jointly trained text-embedding model. Let $c_\theta(y)$ be a model that maps a conditioning input y into a conditioning vector. The LDM loss is then given by:

$$L_{LDM} := \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right], \tag{1}$$

where t is the time step, z_t is the latent noised to time t, ϵ is the unscaled noise sample, and ϵ_θ is the denoising network. Intuitively, the objective here is to correctly remove the noise added to a latent representation of an image. While training, c_θ and ϵ_θ are jointly optimized to minimize the LDM loss. At inference time, a random noise tensor is sampled and iteratively denoised to produce a new image latent, z_0. Finally, this latent code is transformed into an image through the pre-trained decoder $x' = D(z_0)$.

We employ the publicly available 1.4 billion parameter text-to-image model of Rombach et al. (2021), which was pre-trained on the LAION-400M dataset (Schuhmann et al., 2021). Here, c_θ is realized through a BERT (Devlin et al., 2018) text encoder, with y being a text prompt.

We next review the early stages of such a text encoder, and our choice of inversion space.

Text embeddings. Typical text encoder models, such as BERT, begin with a text processing step (Figure 2, left). First, each word or sub-word in an input string is converted to a token, which is an index in some pre-defined dictionary. Each token is then linked to a unique embedding vector that can be retrieved through an index-based lookup. These embedding vectors are typically learned as part of the text encoder c_θ.

In our work, we choose this embedding space as the target for inversion. Specifically, we designate a placeholder string, S_*, to represent the new concept we wish to learn. We intervene in the embedding process and replace the vector associated with the tokenized string with a new, *learned* embedding v_*, in essence "injecting" the concept into our vocabulary. In doing so, we can then compose new sentences containing the concept, just as we would with any other word.

Textual inversion. To find these new embeddings, we use a small set of images (typically 3-5), which depicts our target concept across multiple settings such as varied backgrounds or poses. We find v_* through direct optimization, by minimizing the LDM loss of Equation (1) over images sampled from the small set. To condition the generation, we randomly sample neutral context texts, derived from the CLIP ImageNet templates (Radford et al., 2021). These contain prompts of the form "A photo of S_*", "A rendition of S_*", etc. The full list of templates is provided in the supplementary materials.

Our optimization goal can then be defined as:

$$v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right], \tag{2}$$

and is realized by re-using the same training scheme as the original LDM model, while keeping both c_θ and ϵ_θ fixed. Notably, this is a reconstruction task. As such, we expect it to motivate the learned embedding to capture fine visual details unique to the concept.

Implementation details. Unless otherwise noted, we retain the original hyper-parameter choices of LDM (Rombach et al., 2021). Word embeddings were initialized with the embeddings of a single-word coarse descriptor of the object (*e.g.* "sculpture" and "cat" for the two concepts in Figure 1). Our experiments were conducted using $2\times$V100 GPUs with a batch size of 4. The base learning rate was set to 0.005. Following LDM, we further scale the base learning rate by the number of GPUs and the batch size, for an effective

rate of 0.04. All results were produced using $5,000$ optimization steps. We find that these parameters work well for most cases. However, we note that for some concepts, better results can be achieved with fewer steps or with an increased learning rate.

4 Qualitative comparisons and applications

In the following section, we demonstrate a range of applications enabled through Textual Inversions, and provide visual comparisons to the state-of-the-art and human-captioning baselines.

4.1 Image variations

Figure 3: Object variations generated using our method, the CLIP-based reconstruction of DALLE-2 (Ramesh et al., 2022), and human captions of varying lengths. Our method generates variations which are typically more faithful to the original subject.

We begin by demonstrating our ability to capture and recreate variations of an object using a single pseudo-word. In Figure 3 we compare our method to two baselines: LDM guided by a human caption and DALLE-2 guided by either a human caption or an image prompt. Captions were collected using Mechanical Turk. Annotators were provided with four images of a concept and asked to describe it in a manner that could allow an artist to recreate it. We asked for both a short ($\leq$ 12 words) and a long ($\leq$ 30 words) caption. In total, we collected 10 captions per concept — five short and five long. Figure 3 shows multiple results generated with a randomly chosen caption for each setup. Additional large-scale galleries showing our uncurated reconstructions are provided in the supplementary.

As our results demonstrate, our method better captures the unique details of the concept. Human captioning typically captures the most prominent features of an object, but provides insufficient detail to reconstruct finer features like color patterns (*e.g.* of the teapot). In some cases (*e.g.* the skull mug) the object itself may be exceedingly difficult to describe through natural language. When provided with an image, DALLE-2 is able to recreate more appealing samples, particularly for well-known objects with limited detail (Aladdin's

lamp). However, it still struggles with unique details of personalized objects that the image encoder (CLIP) is unlikely to have seen (mug, teapot). In contrast, our method can successfully capture these finer details, and it does so using only a single word embedding. However, note that while our creations are more similar to the source objects, they are still variations that may differ from the source.

4.2 Text-guided synthesis

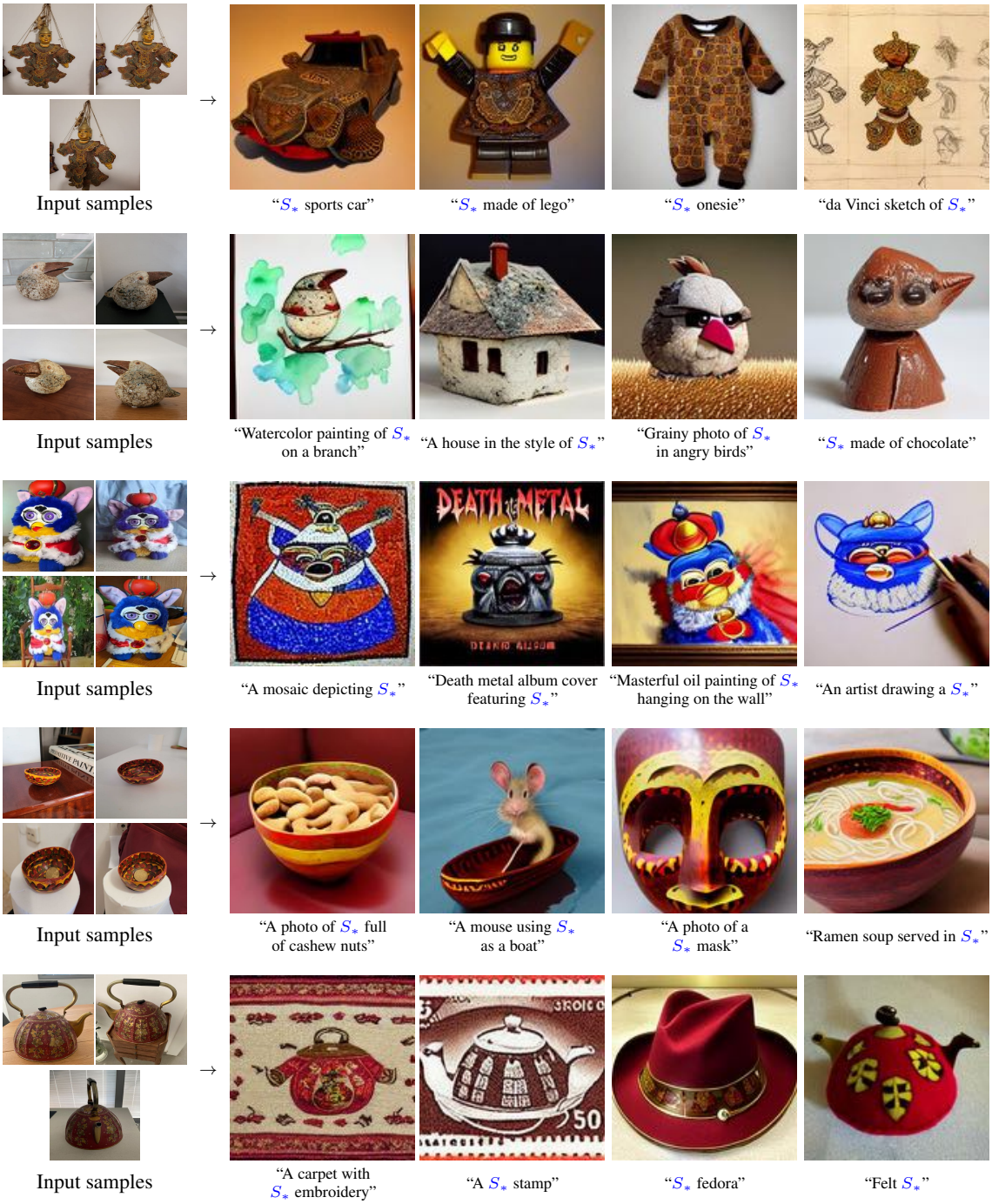


Figure 4: Additional text-guided personalized generation results. In each row, we shows exemplars from the image set representing the concept (left), and novel compositions using the pseudo-word derived from these samples (right).

In Figures 1 and 4 we show our ability to compose novel scenes by incorporating the learned pseudo-words into new conditioning texts. For each concept, we show exemplars from our training set, along with an array of generated images and their conditioning texts. As our results demonstrate, the frozen text-to-image model is able to jointly reason over both the new concepts and its large body of prior knowledge, bringing them together in a new creation. Importantly, despite the fact that our training goal was generative in nature, our pseudo-words still encapsulate semantic concepts that the model can then leverage. For example, observe the bowl's ability (row four) to contain other objects like food, or the ability to preserve the Furby's bird-like head and crown while adapting his palette to better match a prompt (album cover, row three). Additional concepts and texts are provided in the supplementary materials.

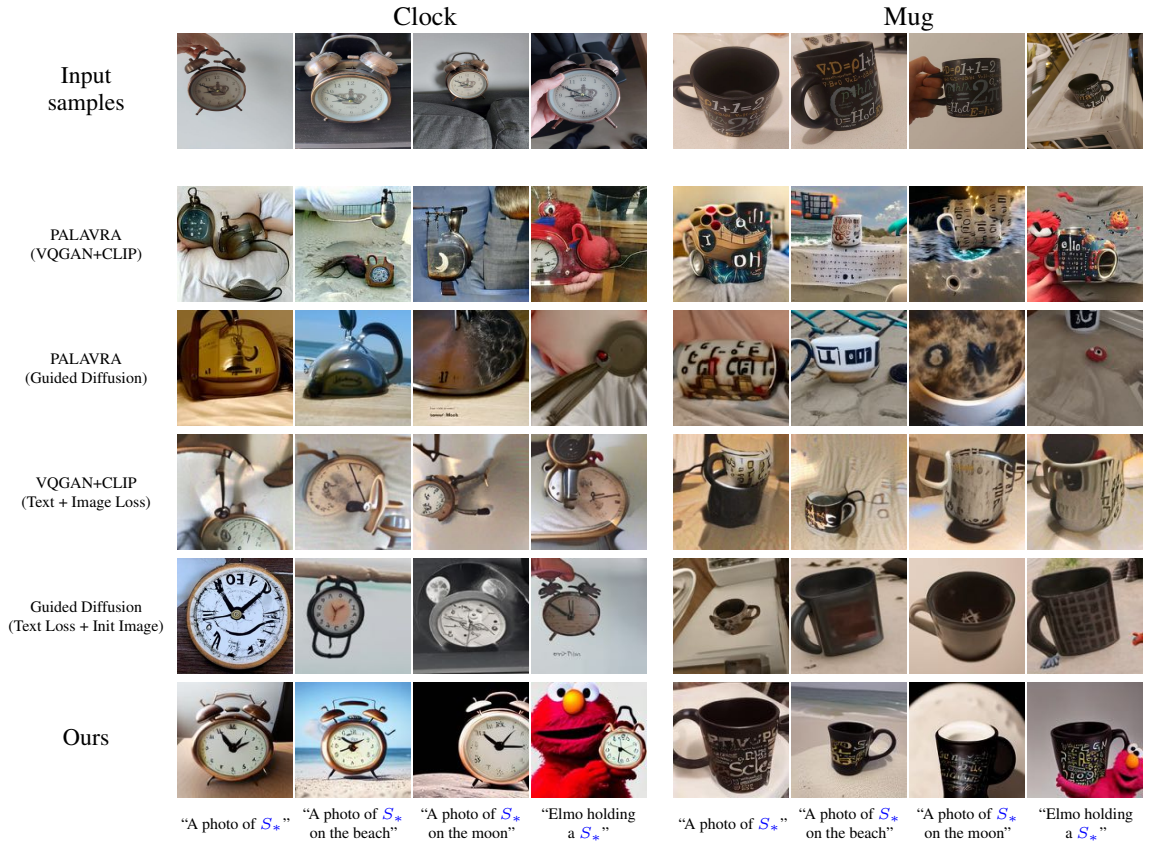


Figure 5: Comparisons to alternative personalized creation approaches. Our model can more accurately preserve the subject, and can reason over both the novel embedding and the rest of the caption.

To better evaluate our ability to compose objects into new scenes, we compare our method to several personalization baselines (Figure 5). In particular, we consider the recent PALAVRA (Cohen et al., 2022), which is most similar to our own work. PALAVRA encodes object sets into the textual embedding space of CLIP, using a mix of contrastive learning and cyclic consistency goals. We find a new pseudo-word using their approach and use it to synthesize new images by leveraging VQGAN-CLIP (Crowson et al., 2022) and CLIP-Guided Diffusion (Crowson, 2021). As a second baseline, we apply the CLIP-guided models of Crowson *et al*. while trying to jointly minimize the CLIP-based distances to both the training set images and to the target text (VQGAN-CLIP) or by initializing the optimization with an input image from our set (Guided Diffusion). For the latter, we chose image-based initializations as we observed that they outperform the use of images in the optimization loss. Similar observations were reported in Disco Diffusion (Letts et al., 2021).

The images produced by PALAVRA (rows 2, 3) typically contain elements from the target prompt (*e.g*. a beach, a moon) but they fail to accurately capture the concept and display considerable visual corruption. This is unsurprising, as PALAVRA was trained with a discriminative goal. In their case, the model needs to only encode enough information to distinguish between two typical concepts (*e.g*. it may be sufficient to remember the mug was black-and-white with text-like symbols). Moreover, their word-discovery process had no need to remain in regions of the embedding space that contain embedding vectors that can be mapped to outputs on the natural image manifold. In the case of the text-and-image guided synthesis methods (rows 4, 5),

results appear more natural and closer to the source image, but they fail to generalize to new texts. Moreover, as our method builds upon pre-trained, large-scale text-to-image synthesis models, we can optimize a single pseudo-word and re-use it for a multitude of new generations. The baseline models, meanwhile, use CLIP for test-time optimization and thus require expensive optimization for every new creation.

4.3 Style transfer

A typical use-case for text-guided synthesis is in artistic circles, where users aim to draw upon the unique style of a specific artist and apply it to new creations. Here, we show that our model can also find pseudo-words representing a specific, unknown style. To find such pseudo-words, we simply provide the model with a small set of images with a shared style, and replace the training texts with prompts of the form: "A painting in the style of S_*". Results are shown in Figure 6. They serve as further demonstration that our ability to capture concepts extends beyond simple object reconstructions and into more abstract ideas.

Note that this differs from traditional style transfer, as we do not necessarily wish to maintain the content of some input image. Instead, we offer the network the freedom to decide how to depict the subject, and merely ask for an appropriate style.

Figure 6: The textual-embedding space can represent more abstract concepts, including styles. This allows us to discover words which can be used for style-guided generation. Image credits: @QinniArt (top), @David Revoy (bottom). Image reproduction authorized for non-commercial use only.

4.4 Concept compositions

In Figure 7 we demonstrate compositional synthesis, where the guiding text contains multiple learned concepts. We observe that the model can concurrently reason over multiple novel pseudo-words at the same time. However, it struggles with relations between them (*e.g.* it fails to place two concepts side-by-side). We hypothesize that this limitation arises because our training considers only single concept scenes, where the concept is at the core of the image. Training on multi-object scenes may alleviate this shortcoming. However, we leave such investigation to future work.

4.5 Bias reduction

A common limitation of text-to-image models is that they inherit the biases found in the internet-scale data used to train them. These biases then manifest in the generated samples. For example, the DALLE-2 system card (Mishkin et al., 2022) reports that their baseline model tends to produce images of people that are white-passing and male-passing when provided with the prompt "A CEO". Similarly, results for "wedding", tend to assume Western wedding traditions, and default to heterosexual couples.

Here, we demonstrate that we can utilize a small, curated dataset in order to learn a new "fairer" word for a biased concept, which can then be used in place of the original to drive a more inclusive generation.

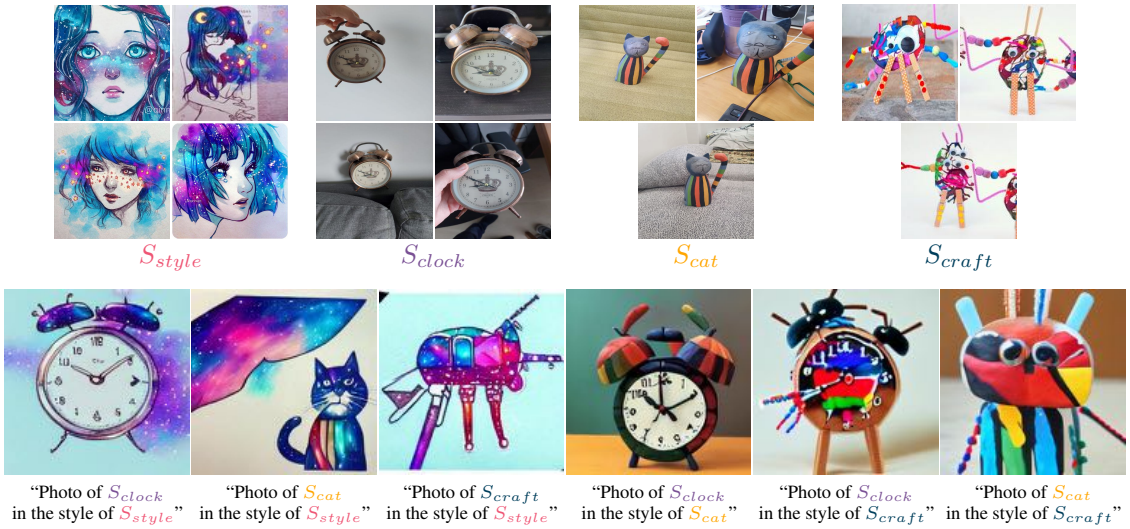


Figure 7: Compositional generation using two learned pseudo-words. The model is able to combine the semantics of two concepts when using a prompt that combines them both. It is limited in its ability to reason over more complex relational prompts, such as placing two concepts side-by-side. Image credits: @QinniArt (left), @Leslie Manlapig (right). Reproductions authorized for non-commercial / non-print use respectively.

Figure 8: Bias Reduction. Uncurated samples synthesized with pretrained biased embeddings (left) and our *de*biased embeddings (right). Our approach can be used to reduce bias by learning new pseudo-words for known concepts. These can be optimized using small datasets, which can be carefully curated for diversity.

Specifically, in Figure 8 we highlight the bias encoded in the word "Doctor", and show that this bias can be reduced (*i.e.* we increase perceived gender and ethnic diversity) by learning a new embedding from a small, more diverse set.

4.6 Downstream applications

Finally, we demonstrate that our pseudo-words can be used in downstream models that build on the same initial LDM model. Specifically, we consider the recent Blended Latent Diffusion (Avrahami et al., 2022a) which enables localized text-based editing of images via a mask-based blending process in the latent space of an LDM. In Figure 9 we demonstrate that this localized synthesis process can also be conditioned on our learned pseudo-words, without requiring any additional modifications of the original model.

4.7 Image curation

Unless otherwise noted, results in this section are partially curated. For each prompt, we generated 16 candidates (or six for DALLE-2) and manually selected the best result. We note that similar curation processes with larger batches are typically employed in text-conditioned generation works (Avrahami et al., 2022b; Ramesh et al., 2021; Yu et al., 2022), and that one can automate this selection process by using CLIP to

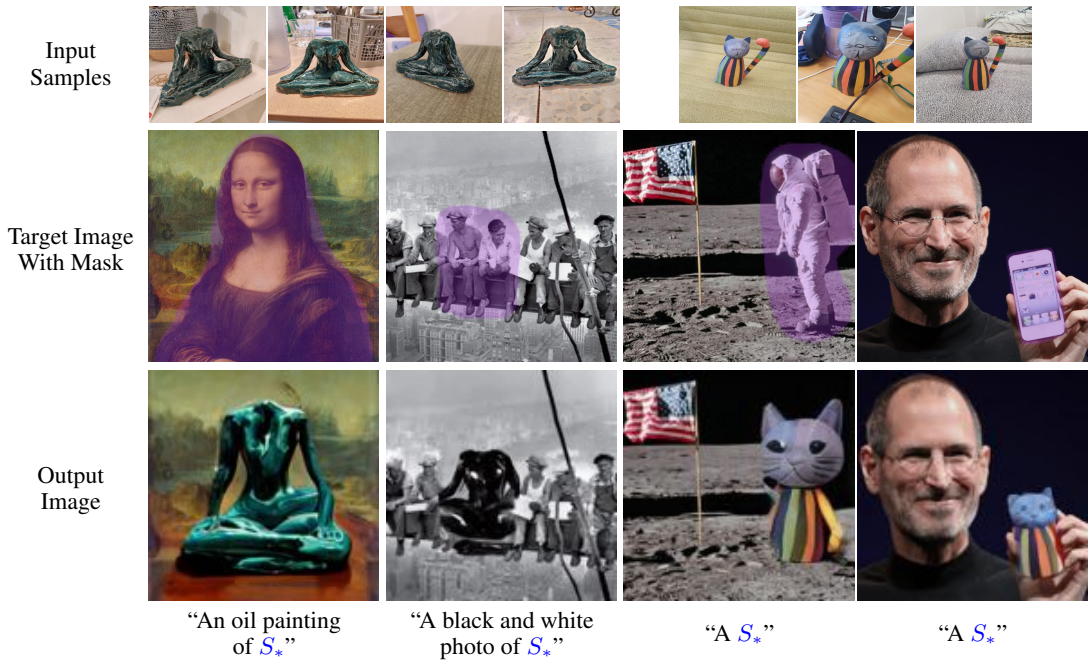


Figure 9: Our words can be used with downstream models that build on LDM. Here, we perform localized image editing using Blended Latent Diffusion (Avrahami et al., 2022a)

rank images. In the supplementary materials, we provide large-scale, uncurated galleries of generated results, including failure cases.

5 Quantitative analysis

Inversion into an uncharted latent space provides us with a wide range of possible design choices. Here, we examine these choices in light of the GAN inversion literature and discover that many core premises (such as a distortion-editability tradeoff (Tov et al., 2021; Zhu et al., 2020b)) also exist in the textual embedding space. However, our analysis reveals that many of the solutions typically used in GAN inversion fail to generalize to this space, and are often unhelpful or actively harmful.

5.1 Evaluation metrics

To analyze the quality of latent space embeddings, we consider two fronts: reconstruction and editability. First, we wish to gauge our ability to replicate the target concept. As our method produces variations on the concept and not a specific image, we measure similarity by considering semantic CLIP-space distances. Specifically, for each concept, we generate a 64 of images using the prompt: "A photo of S_*". Our reconstruction score is then the average pair-wise CLIP-space cosine-similarity between the generated images and the images of the concept-specific training set.

Second, we want to evaluate our ability to modify the concepts using textual prompts. To this end, we produce a set of images using prompts of varying difficulty and settings. These range from background modifications ("A photo of S_* on the moon"), to style changes ("An oil painting of S_*"), and a compositional prompt ("Elmo holding a S_*").

For each prompt, we synthesize 64 samples using 50 DDIM steps, calculate the average CLIP-space embedding of the samples, and compute their cosine similarity with the CLIP-space embedding of the textual prompts, where we omit the placeholder S_* (*i.e.* "A photo of on the moon"). Here, a higher score indicates better editing capability and more faithfulness to the prompt itself. Note that our method does not involve the direct optimization of the CLIP-based objective score and, as such, is not sensitive to the adversarial scoring flaws outlined by Nichol et al. (2021).

5.2 Evaluation setups

We evaluate the embedding space using a set of experimental setups inspired by GAN inversion:

Extended latent spaces Following Abdal et al. (2019), we consider an extended, multi-vector latent space. In this space, S_* is embedded into multiple learned embeddings, an approach that is equivalent to describing the concept through multiple learned pseudo-words. We consider an extension to two and three pseudo-words (denoted $2-word$ and $3-word$, respectively). This setup aims to alleviate the potential bottleneck of a single embedding vector to enable more accurate reconstructions.

Progressive extensions We follow Tov et al. (2021) and consider a progressive multi-vector setup. Here, we begin training with a single embedding vector, introduce a second vector following $2,000$ training steps, and a third vector after $4,000$ steps. In this scenario, we expect the network to focus on the core details first, and then leverage the additional pseudo-words to capture finer details.

Regularization Tov et al. (2021) observed that latent codes in the space of a GAN have increased editability when they lie closer to the code distribution which was observed during training. Here, we investigate a similar scenario by introducing a regularization term that aims to keep the learned embedding close to existing words. In practice, we minimize the L2 distance of the learned embedding to the embedding of a coarse descriptor of the object (*e.g.* "sculpture" and "cat" for the images in Figure 1).

Per-image tokens Moving beyond GAN-based approaches, we investigate a novel scheme where we introduce unique, per-image tokens into our inversion approach. Let $\{x_i\}_{i=1}^n$ be the set of input images. Rather than optimizing a single word vector shared across all images, we introduce both a universal placeholder, S_*, and an additional placeholder unique to each image, $\{S_i\}_{i=1}^n$, associated with a unique embedding v_i. We then compose sentences of the form "A photo of S_* with S_i", where every image is matched to sentences containing its own, unique string. We jointly optimize over both S_* and $\{S_i\}_{i=1}^n$, using Equation (2). The intuition here is that the model should prefer to encode the shared information (*i.e.* the concept) in the shared code S_* while relegating per-image details such as the background to S_i.

Human captions In addition to the learned-embedding setups, we compare to human-level performance using the captions outlined in Section 4.1. Here, we simply replace the placeholder strings S_* with the human captions, using both the short and long-caption setups.

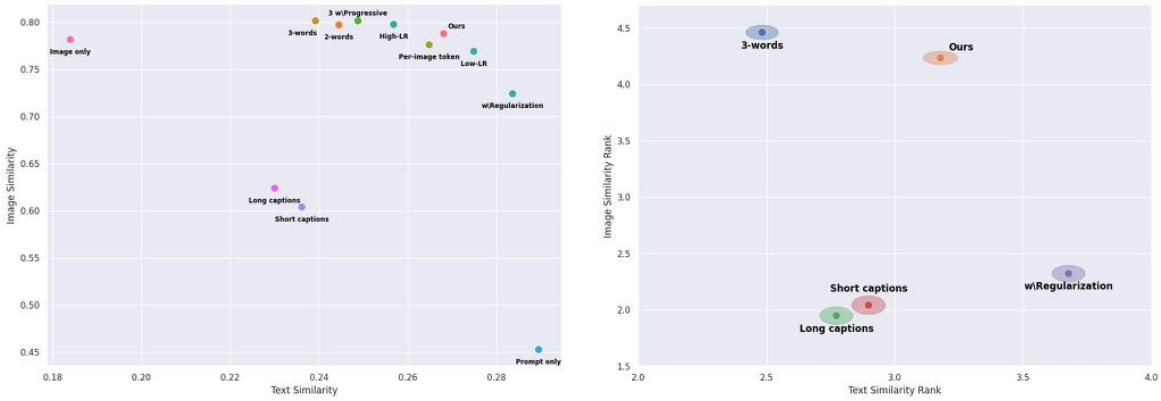
Reference setups To provide intuition for the scale of the results, we add two reference baselines. First, we consider the expected behavior from a model that always produces copies of the training set, regardless of the prompt. For that, we simply use the training set itself as the "generated sample". Second, we consider a model that always aligns with the text prompt but ignores the personalized concept. We do so by synthesizing images using the evaluation prompts but without the pseudo-word. We denote these setups as "Image Only" and "Prompt Only", respectively.

Textual-Inversion Finally, we consider our own setup, as outlined in Section 3. We further evaluate our model with an increased learning rate ($2e-2$, "High-LR") and a decreased learning rate ($1e-4$, "Low-LR").

Additional setups In the supplementary, we consider two additional setups for inversion: a pivotal tuning approach (Roich et al., 2021; Bau et al., 2020), where the model itself is optimized to improve reconstruction, and DALLE-2 (Ramesh et al., 2022)'s bipartite inversion process. We further analyze the effect of the image-set size on reconstruction and editability.

5.3 Results

Our evaluation results are summarized in Figure 10(a). We highlight four observations of particular interest: First, the semantic reconstruction quality of our method and many of the baselines is comparable to simply sampling random images from the training set. Second, the single-word method achieves comparable reconstruction quality, and considerably improved editability over all multi-word baselines. These points outline the impressive flexibility of the textual embedding space, showing that it can serve to capture new concepts with a high degree of accuracy while using only a single pseudo-word.



(a) (b)

Figure 10: Qualitative evaluation results. (a) CLIP-based evaluations. The single-word model (ours) represents an appealing point on the distortion-editability curve, and can be moved along it by changing the learning rate. (b) User study results. These results portray a similar distortion-editability curve, and demonstrate that the CLIP-based results align with human preference. User study error bars are 95% confidence intervals.

Third, we observe that our baselines outline a distortion-editability trade-off curve, where embeddings that lie closer to the true word distribution (*e.g.* due to regularization, fewer pseudo-words, or a lower learning rate) can be more easily modified, but fail to capture the details of the target. In contrast, deviating far from the word distribution enables improved reconstruction at the cost of severely diminished editing capabilities. Notably, our single-embedding model can be moved along this curve by simply changing the learning rate, offering a user a degree of control over this trade-off.

As a fourth observation, we note that the use of human descriptions for the concepts not only fails to capture their likeness, but also leads to diminished editability. We hypothesize that this is tied to the selective-similarity property outlined in Paiss et al. (2022), where vision-and-language models tend to focus on a subset of the semantically meaningful tokens. By using long captions, we increase the chance of the model ignoring our desired setting, focusing only on the object description itself. Our model, meanwhile, uses only a single token and thus minimizes this risk.

Finally, we note that while our reconstruction scores are on par with those of randomly sampled, real images, these results should be taken with a grain of salt. Our metrics compare *semantic* similarity using CLIP, which is less sensitive to shape-preservation. On this front, there remains more to be done.

5.4 Human evaluations

We further evaluate our models using a user study. Here, we created two questionnaires. In the first, users were provided with four images from a concept's training set, and asked to rank the results produced by five models according to their similarity to these images. In the second questionnaire, users were provided with a text describing an image context ("A photo on the beach") and asked to rank the results produced by the same models according to their similarity to the text.

We used the same target concepts and prompts as the CLIP-based evaluation and collected a total of 600 responses to each questionnaire, for a total of $1,200$ responses. Results are shown in Figure 10(b).

The user-study results align with the CLIP-based metrics and demonstrate a similar reconstruction-editability tradeoff. Moreover, they outline the same limitations of human-based captioning when attempting to reproduce a concept, as well as when editing it.

6 Limitations

While our method offers increased freedom, it may still struggle with learning precise shapes, instead incorporating the "semantic" essence of a concept. For artistic creations, this is often enough. In the future, we

hope to achieve better control over the accuracy of the reconstructed concepts, enabling users to leverage our method for tasks that require greater precision.

Another limitation of our approach is in the lengthy optimization times. Using our setup, learning a single concept requires roughly two hours. These times could likely be shortened by training an encoder to directly map a set of images to their textual embedding. We aim to explore this line of work in the future.

7 Social impact

Text-to-image models can be used to generate misleading content and promote disinformation. Personalized creation could allow a user to forge more convincing images of non-public individuals. However, our model does not currently preserve identity to the extent where this is a concern.

These models are further susceptible to the biases found in the training data. Examples include gender biases when portraying "doctors" and "nurses", racial biases when requesting images of scientists, and more subtle biases such as an over-representation of heterosexual couples and western traditions when prompting for a "wedding" (Mishkin et al., 2022). As we build on such models, our own work may similarly exhibit biases. However, as demonstrated in Figure 8, our ability to more precisely describe specific concepts can also serve as a means for reducing these biases.

Finally, the ability to learn artistic styles may be misused for copyright infringement. Rather than paying an artist for their work, a user could train on their images without consent, and produce images in a similar style. While generated artwork is still easy to identify, in the future such infringement could be difficult to detect or legally pursue. However, we hope that such shortcomings are offset by the new opportunities that these tools could offer an artist, such as the ability to license out their unique style, or the ability to quickly create early prototypes for new work.

8 Conclusions

We introduced the task of personalized, language-guided generation, where a text-to-image model is leveraged to create images of specific concepts in novel settings and scenes. Our approach, "Textual Inversions", operates by *inverting* the concepts into new pseudo-words within the textual embedding space of a pre-trained text-to-image model. These pseudo-words can be injected into new scenes using simple natural language descriptions, allowing for simple and intuitive modifications. In a sense, our method allows a user to leverage multi-modal information — using a text-driven interface for ease of editing, but providing visual cues when approaching the limits of natural language.

Our approach was implemented over LDM (Rombach et al., 2021), the largest publicly available text-to-image model. However, it does not rely on any architectural details unique to their approach. As such, we believe Textual Inversions to be easily applicable to additional, larger-scale text-to-image models. There, text-to-image alignment, shape preservation, and image generation fidelity may be further improved.

We hope our approach paves the way for future personalized generation works. These could be core to a multitude of downstream applications, from providing artistic inspiration to product design.

Acknowledgments We thank Yael Vinker, Roni Paiss and Haggai Maron for reviewing early drafts and helpful suggestions. Tom Bagshaw for discussions regarding artist rights and social impacts, and Omri Avrahami for providing us with early access to Blended Latent Diffusion. This work was partially supported by Len Blavatnik and the Blavatnik family foundation, BSF (grant 2020280) and ISF (grants 2492/20 and 3441/21).

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019.

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8296–8305, 2020.

Rameen Abdal, Peihao Zhu, John Femiani, Niloy J Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. *arXiv preprint arXiv:2112.05219*, 2021.

Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learned compression of images and neural networks. *arXiv preprint arXiv:1704.00648*, 3, 2017.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H. Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing, 2021.

Fernando Amat, Ashok Chandrashekar, Tony Jebara, and Justin Basilico. Artwork personalization at netflix. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pp. 487–488, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3241729. URL https://doi.org/10.1145/3240323.3241729.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022a.

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*, pp. 18208–18218, 2022b.

Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. 38(4), 2019. ISSN 0730-0301. doi: 10.1145/3306346.3323023. URL https://doi.org/10.1145/3306346.3323023.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.

Soulef Benhamdi, Abdesselam Babouri, and Raja Chiky. Personalized recommender system for e-learning environment. *Education and Information Technologies*, 22(4):1455–1477, 2017.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shunsuke Saito, Stephen Lombardi, Shih-en Wei, Danielle Belko, Shoou-i Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 2022.

Yoon Ho Cho, Jae Kyeong Kim, and Soung Hie Kim. A personalized recommender system based on web usage mining and decision tree induction. *Expert systems with Applications*, 23(3):329–342, 2002.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European Conference on Computer Vision (ECCV)*, 2022.

Katherine Crowson. CLIP guided diffusion HQ 256x256. https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj, 2021.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Yuxuan Ding, Lingqiao Liu, Chunna Tian, Jingyuan Yang, and Haoxuan Ding. Don't stop learning: Towards continual learning for the clip model. *arXiv e-prints*, pp. arXiv–2207, 2022.

Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11389–11398, 2022.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021.

Adam Letts, Chris Scalf, Alex Spirin, Tom Mason, Chris Allen, Max Ingham, Mike Howles, Nate Baer, and David Sin. Disco diffusion. https://github.com/alembics/disco-diffusion, 2021.

Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Xing Fan, Edward Guo, and Yang Liu. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In *NAACL 2022*, 2022. URL https://www.amazon.science/publications/overcoming-catastrophic-forgetting-during-domain-adaptation-of-seq2seq-language-generation.

Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: An arbitrary artist-aware image style transfer. *arXiv preprint arXiv:2202.13562*, 2022.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Ana Belen Barragans Martinez, Jose J Pazos Arias, Ana Fernandez Vilas, Jorge Garcia Duque, and Martin Lopez Nores. What's on tv tonight? an efficient and effective personalized recommender system of tv programs. *IEEE Transactions on Consumer Electronics*, 55(1):286–294, 2009.

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. *arXiv preprint arXiv:2112.03221*, 2021.

Pamela Mishkin, Lama Ahmad, Miles Brundage, Gretchen Krueger, and Girish Sastry. Dall·e 2 preview - risks and limitations. 2022.

Ryan Murdock. The big sleep, 2021. https://twitter.com/advadnoun/status/1351038053033406468.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *arXiv preprint arXiv:2203.17272*, 2022.

Roni Paiss, Hila Chefer, and Lior Wolf. No token left behind: Explainability-aided image classification and generation. *arXiv preprint arXiv:2204.04908*, 2022.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv preprint arXiv:2103.17249*, 2021.

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14104–14113, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

N. Rathvon. *Early Reading Assessment: A Practitioner's Handbook*. Guilford Publications, 2004. ISBN 9781572309845. URL https://books.google.co.uk/books?id=zX8YzwEACAAJ.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pp. 9489–9502. PMLR, 2021.

Gabriel Skantze and Bram Willemsen. Collie: Continual learning of language grounding from language-image embeddings. *Journal of Artificial Intelligence Research*, 74:1201–1223, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *arXiv preprint arXiv:2203.08063*, 2022.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212, 2021.

Rotem Tzaban, Ron Mokady, Rinon Gal, Amit H. Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey, 2021.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.

Paul Yacoubian. Avocado bag, 2022. https://twitter.com/PaulYacoubian/status/1542867718071779330.

Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020a.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pp. 597–613. Springer, 2016.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5810, 2019.

Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Improved stylegan embedding: Where are the good latents?, 2020b.

Supplementary Materials
An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Rinon Gal[1,2] **Yuval Alaluf**[1] **Yuval Atzmon**[2] **Or Patashnik**[1]

Amit H. Bermano[1] **Gal Chechik**[2] **Daniel Cohen-Or**[1]

[1]**Tel-Aviv University** [2]**NVIDIA**

A Additional inversion approaches

In addition to the setups outlined in the core paper, we investigated two recent approaches to inversion: Bipartite DDIM-inversion (Ramesh et al., 2022; Dhariwal & Nichol, 2021) and pivotal tuning (Roich et al., 2021). Below we outline both methods and our experimental results.

Bipartite inversion Dhariwal & Nichol (2021) demonstrated that the DDIM sampling (Song et al., 2020) process can be inverted through a closed-form iterative approach. Specifically, their approach can find a latent noise vector x_T which will be denoised into a specific target image when the denoising process is conditioned on a given code $c_\theta(y)$. In (Ramesh et al., 2022), they further demonstrate that when the conditioning code is an output of CLIP, one can later modify this code using text-derived directions in CLIP's multi-modal embedding space, while keeping the initial noise, x_T, fixed. This induces semantic changes in the image while maintaining the general structure of the original object.

Here, we investigate a similar approach. However, rather than modifying the conditioning code $c_\theta(y)$ directly, we change the conditioning text y. Specifically, we first find an appropriate pseudo-word for our target concept. Then, we find x_T for a given image of the concept using the text "A photo of S_*" and the closed-form solution of Dhariwal & Nichol (2021). Finally, we modify the conditioning text but keep x_T frozen. The results are shown in Figure 11 (left). Here, we observe that when using LDM's typical guidance (Ho & Salimans, 2021) scales (5-10), the denoiser network is unable to maintain the original object's structure through prompt changes. When reducing the guidance scale, the outline of the original image becomes visible. However, alignment with the prompt is poor.

Such guidance-dependent structure drift has also been demonstrated for GLIDE (Nichol et al., 2021). However, this effect is reduced in DALL-E2 (Ramesh et al., 2022) (their Figure 9). Notably, state-of-the-art models (Saharia et al., 2022; Ramesh et al., 2022) typically employ guidance scales (~ 2) which are significantly lower than LDM's — within the range where we observe structure preservation, but no prompt-matching. This gives us hope that a bipartite inversion would allow better shape preservation in more powerful generative models.

Pivotal Tuning In the field of GAN inversion, it has been shown (Roich et al., 2021; Bau et al., 2019) that one may largely avoid the reconstruction-editability tradeoff using a two-stage optimization process. First, an image is inverted into "pivot" code in a well-behaved region of the latent space, using standard optimization. This typically results in a highly editable code, but with poor identity preservation. As a second step, the generator is fine-tuned so that the first step's pivot code will more accurately reproduce the inverted image. It was further demonstrated that such localized tuning can maintain the appealing properties of the latent space and retain similar latent-editing capabilities.

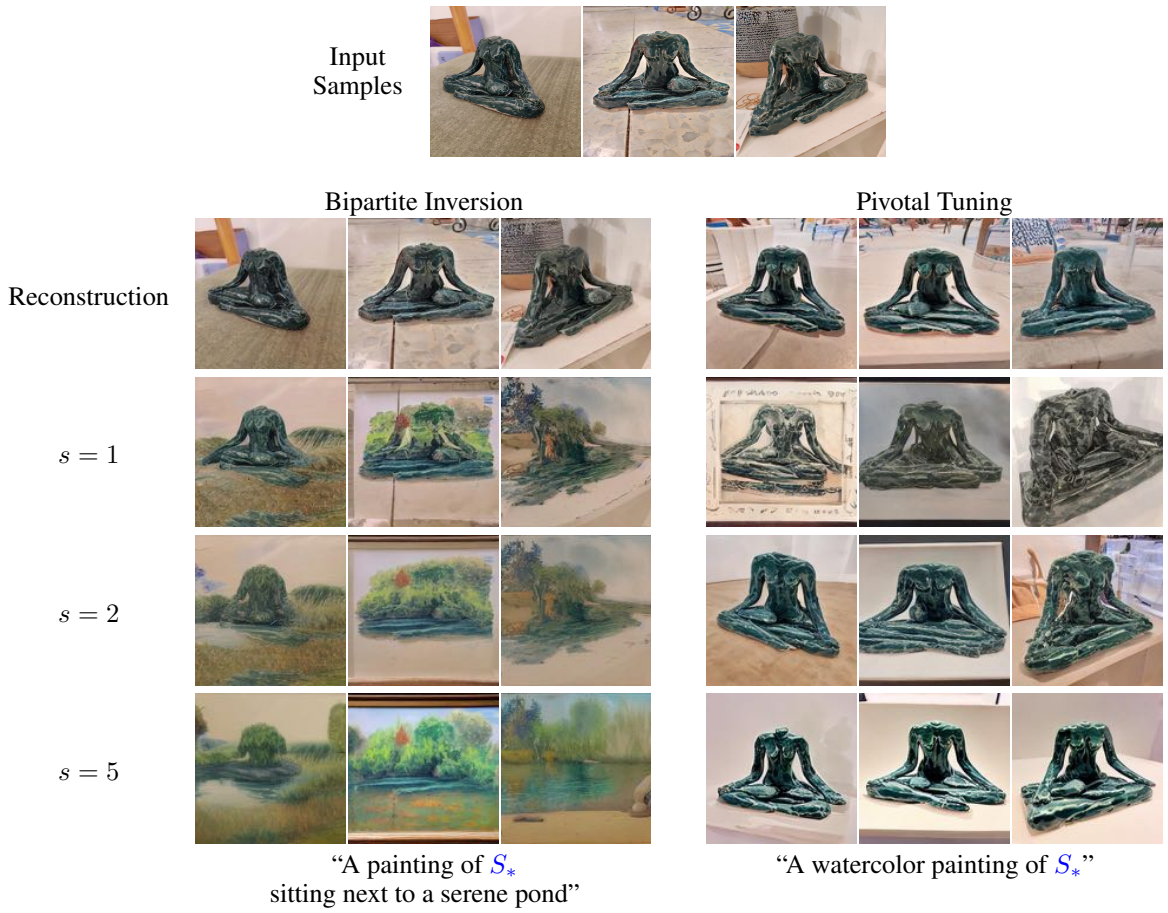


Figure 11: Advanced inversion results using Bipartite Inversion (Ramesh et al., 2022) (left) and Pivotal Tuning (Roich et al., 2021) (right). s is the guidance scale. Reconstructions were obtained using the prompt "A photo of S_*". Bipartite inversion allows for more accurate reconstructions without modifying the model, but their structure is lost for complex prompts in high guidance scales. Pivotal tuning improves shapes at the cost of visual artifacts, and fail to adhere to simple prompts at high guidance scales. Note that the pivotal tuning results use a random noise input, while the bipartite results use a fixed noise for each column.

Here, we investigate a similar approach in order to improve reconstruction. We first optimize a pseudo-word using our baseline method. Then, we fine-tune the generator such that sentences of the form "A photo of S_*" will better reconstruct the concept-specific training set images.

Our initial investigation reveals that naïve applications of this approach lead to improved shape preservation, but also to a severe collapse of editing at high guidance scales. See Figure 11 (right) for examples.

However, a more involved application of this same principle (*e.g.* by combining it with a similar process to the bipartite-inversion outlined below, or by tuning around results produced with higher guidance scales) might overcome these issues. We leave such investigation to future work.

B Effect of training set size

We investigated the effect of the concept's training set size on the results. Specifically, we consider the headless sculpture object of Figure 1 (top row). We inverted the object using our standard model but sweeped over dataset sizes ranging from a single image to 25 samples. For ease of comparison, we further report the image-only, prompt-only, and human caption based scores for the same single object. The results are shown in Figure 12.

Using additional images leads to optimized embeddings which reside farther away from real word embeddings, harming editability. Our method operates best when provided with 5 images.

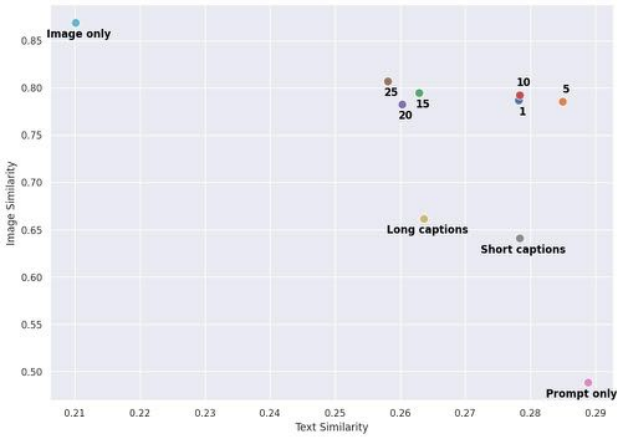


Figure 12: Quantitative evaluation of the effects of the training set size. Significant increases to dataset sizes leads to larger deviation from the real-word distribution. This impacts editability and offers paltry improvement in reconstruction. Our approach shows the best results with ~ 5 images.

C Additional results

We provide additional results of personalized generation using our method. In Figure 13 we show additional text-guided synthesis results.

In Figure 14 we show large-scale galleries of uncurated results generated with the prompt "A photo of S_*". In Figures 15 and 16 we provide large-scale galleries of uncurated results generated with a wide assortment of prompts. These are intended to provide a sense of the quality of images produced and cherry-picking involved when generating the samples in the core paper. Note that these results also contain demonstrations of typical failure cases, such as difficult relational prompts (Figure 15, rows 2, 5).

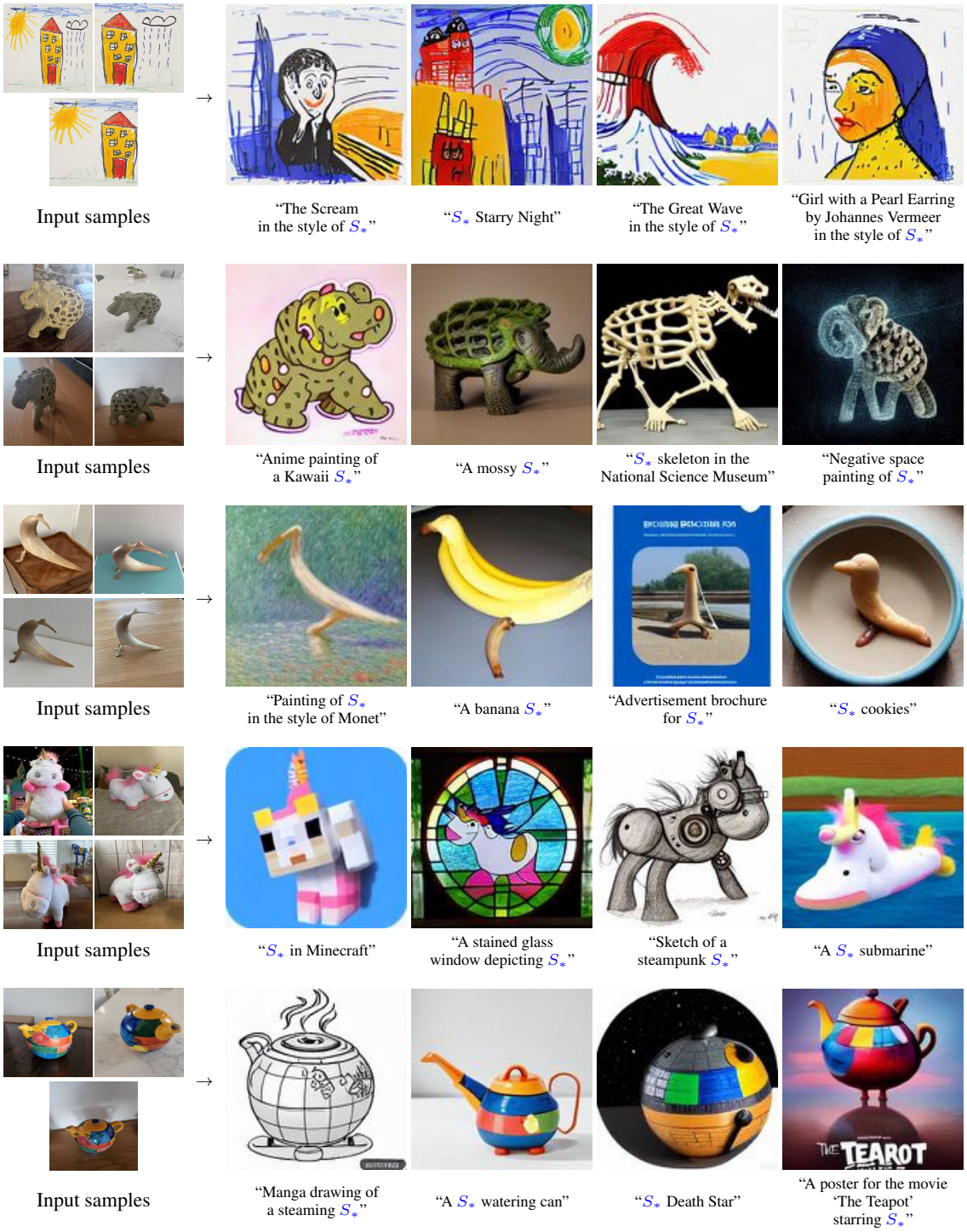


Figure 13: Injecting user-specific concepts into new scenes. Our method can change a concept's style, composition, or use it to inspire new creations. Top row image credits: @Øyvind Holmstad.

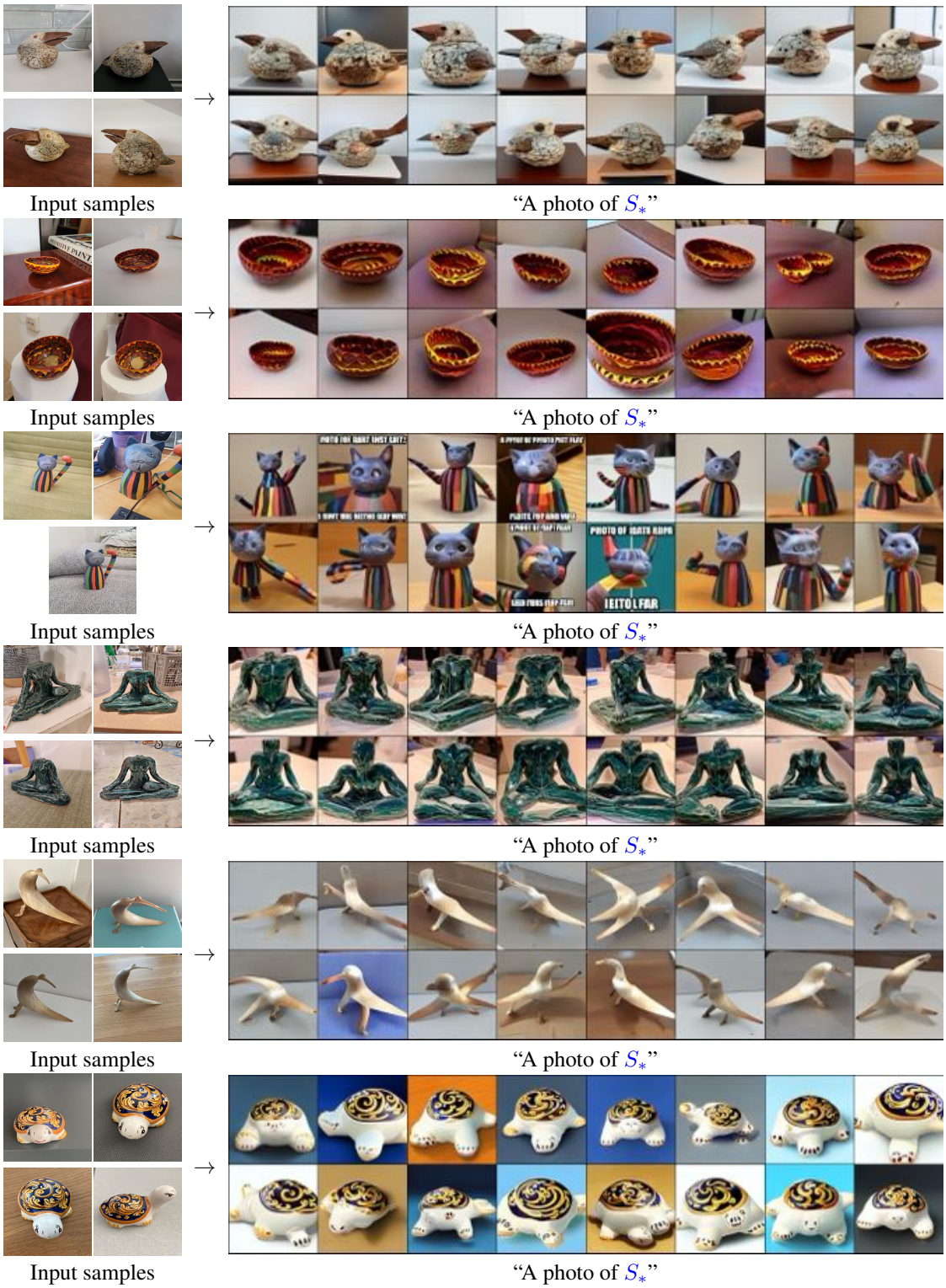
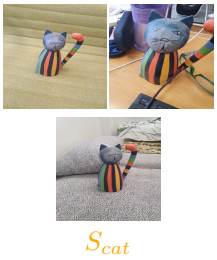


Figure 14: Uncurated samples of object variations created using the prompt ”A photo of S_*”.



S_{cat}

"Banksy art of S_{cat}"

"Painting of a S_{cat} riding a dragon"

S_{clock}

"Painting of two S_{cat} fishing on a boat"

"A S_{cat} themed lunchbox"

S_{bird}

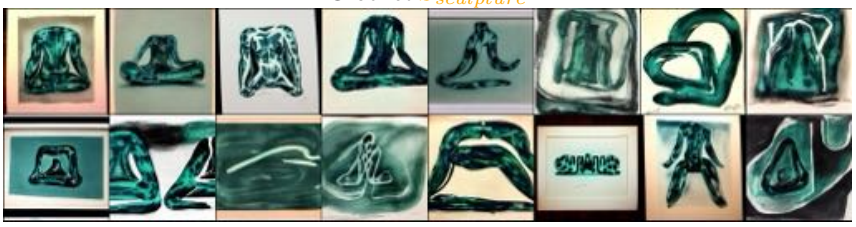
"A painting of S_{cat} holding a S_{clock}"

"Ukiyo-e painting of S_{bird}"

Figure 15: Uncurated samples generated with context prompts. Quality and prompt-matching varies within the sample. However, we observe that a batch size of 16 is typically sufficient to ensure several good samples.

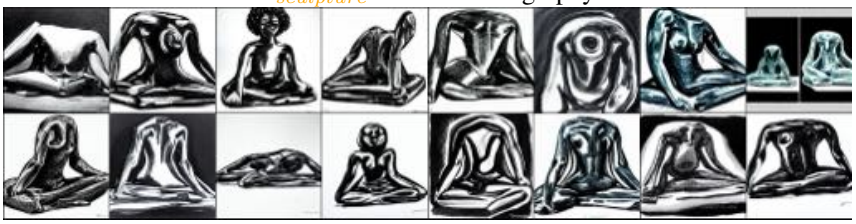
$S_{sculpture}$

"Crochet $S_{sculpture}$"



"$S_{sculpture}$ ink wash calligraphy"

S_{craft}



"A black and white sketch of $S_{sculpture}$"

"A S_{craft} backpack"

S_{style}

"A painting of S_{craft} invading from space"

"A waterfall in the style of S_{style}"

Figure 16: Additional uncurated samples generated with context prompts. Quality and prompt-matching varies within the sample. However, we observe that a batch size of 16 is typically sufficient to ensure several good samples. Image credits: @QinniArt (bottom), authorized for non-commercial use only.

D Training prompt templates

Below we provide the list of text templates used when optimizing a pseudo-word:

- "a photo of a S_*.",
- "a rendering of a S_*.",
- "a cropped photo of the S_*.",
- "the photo of a S_*.",
- "a photo of a clean S_*.",
- "a photo of a dirty S_*.",
- "a dark photo of the S_*.",
- "a photo of my S_*.",
- "a photo of the cool S_*.",
- "a close-up photo of a S_*.",
- "a bright photo of the S_*.",
- "a cropped photo of a S_*.",
- "a photo of the S_*.",
- "a good photo of the S_*.",
- "a photo of one S_*.",
- "a close-up photo of the S_*.",
- "a rendition of the S_*.",
- "a photo of the clean S_*.",
- "a rendition of a S_*.",
- "a photo of a nice S_*.",
- "a good photo of a S_*.",
- "a photo of the nice S_*.",
- "a photo of the small S_*.",
- "a photo of the weird S_*.",
- "a photo of the large S_*.",
- "a photo of a cool S_*.",
- "a photo of a small S_*.",