A Comprehensive Study On Custom Image Generation Using Textual Inversion

Benjamín Farías V.
Pontificia Universidad Católica de Chile
bffarias@uc.cl

Abstract

Large scale 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 novel concepts, modify them, or manipulate them in some other fashion. In this work, we extensively study a method that is able to achieve these tasks, called *Textual Inversion*. Using only a few images for training, this approach is able to learn these concepts in a very efficient manner, leveraging the vast knowledge that the base models contain. Through a set of experiments, both qualitative and quantitative, we show that this method is perfectly applicable in real scenarios, specially in the context of digital art creation. Finally, we review a multitude of examples of generated images, discussing interesting aspects that can be observed from them.

1 Introduction

Recently, large scale text-to-image models, such as *Latent Diffusion Models* (*LDM*) [**3**], have proved to be really capable when it comes to reasoning over natural language descriptions and producing novel images conditioned on said descriptions. They allow users to synthesize novel scenes with unseen compositions and produce high quality images in a myriad of artistic styles. These tools have been successfully used in tasks such as artistic creation, product design, and as sources of inspiration for all sorts of fields. However, this approach is limited by the user's ability to express the desired target's visual semantics through text.

In this work, we study a method named *Textual Inversion* [**1**], which overcomes this limitation by **finding** new words in the textual embedding space of these pre-trained text-to-image models. This is done via the introduction of new pseudo-word tokens into the text encoder layer of the model, accompanied by a new embedding that the encoder has to learn. The learning process is simply performed as a standard weight optimization, just like any other fine-tuning approach. The conducted experiments support the hypothesis that this method is very effective, as well as efficient, for all considered tasks of interest.

In summary, this paper is organized as follows. **Section 2** describes previous work and gives background information about *LDM* models, as well as *fine-tuning* approaches that work on top of them. **Section 3** presents the *Textual Inversion* method, the main ideas behind it, and the setup for all experiments that will be reviewed later in **Sections 4, 5 and 6**. Finally, **Section 7** presents the main conclusions and future research, relative to this work.

2 Related Work

2.1 Latent Diffusion Models

Text-guided image synthesis has been extensively studied in the context of adversarial networks, leveraging attention mechanisms or cross-modal contrastive approaches. More recently, impressive visual results were achieved by using what is known as *Latent Diffusion Models* (*LDM*).

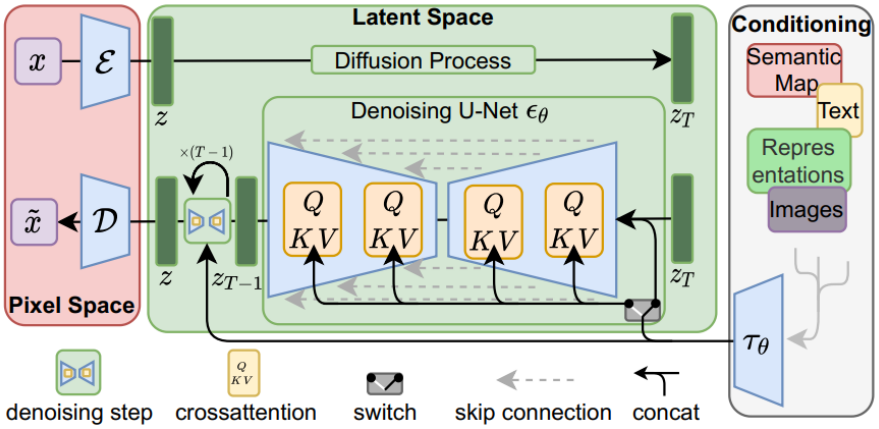

Figure 1: Outline of the *Stable Diffusion* process.

LDM work by decomposing the image formation process into a sequential application of denoising *autoencoders*, which work in a latent space of low dimensional vectors [**3**]. One such model that is currently very popular is known as *Stable Diffusion*.

The *Stable Diffusion* model (**Figure 1**) works by compressing the information from both the training images and the conditioning textual prompts into a low dimensional latent space, which can be done using *autoencoders*. Then, it learns to generate image representations close to those of the target images, via a process of noise reduction (*diffusion*). This *diffusion* process is repeated a fixed amount of times, using *autoencoders* to continually try to get close to the target embedding. Additionally, this search is also guided by a conditional prompt, in this case, a textual prompt that is transformed into an embedding in the latent space with the help of a pre-trained *CLIP* model's text encoder [**2**]. Finally, once the predicted embedding is obtained, an *autodecoder* maps it back into an image in pixel space, which is the output.

The method that we study, *Textual Inversion* (*TI*), builds on these open-ended, conditional synthesis models. Rather than training a new model from scratch, *TI* functions as a *fine-tuning* strategy, maintaining almost the entire model frozen.

2.2 DreamBooth

Currently, one of the most powerful *fine-tuning* methods for *LDM* models is known as *DreamBooth* [**4**]. This method aims to expand the language-vision vocabulary of the large scale models, learning new textual and visual representations for novel subjects that the user wants to generate.

This approach is powerful, but requires training parameters in certain layers of both the textual and visual-oriented components of the network, which can be slow (a few hours of training).

2.3 Textual Inversion

This approach, as mentioned previously, is similar to *DreamBooth*, but more efficient, since it only trains on the embedding layer of the text encoder [**1**] (a few minutes to a couple hours of training). Due to this characteristics, we chose this method to perform experiments and studies, which are presented later on in this work.

3 Method

The *Textual Inversion* method (**Figure 2**) works by adding a new special token to the embedding layer of the pre-trained text encoder (*CLIP*) [**2**]. Then, it directly optimizes these embeddings, by minimizing the *MSE* loss over images sampled from the training set. To condition the generation, we use prompts that generically describe the concept represented by the special token. Throughout this entire process, all layers and parameters of the pre-trained model are frozen, except for the specific embedding of the new special token that was added [**1**].

The initial values for the special token's embedding come from a manually selected initializer

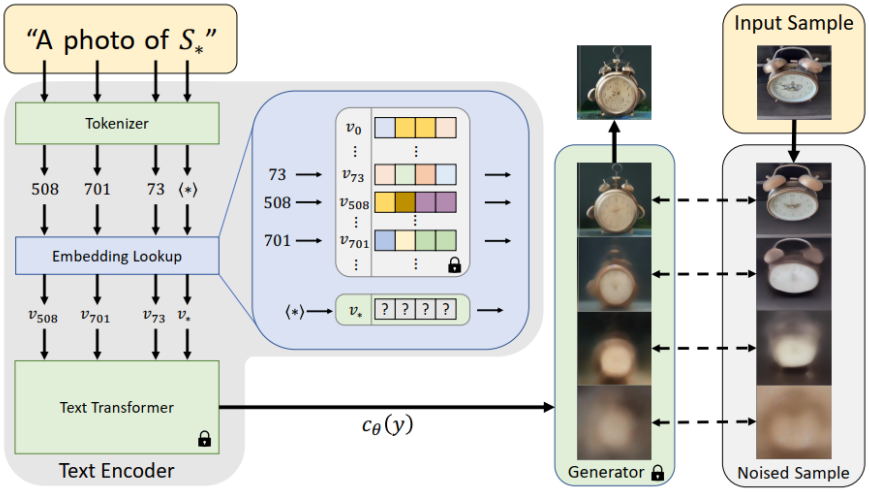

Figure 2: Outline of the *Textual Inversion* process.

concept, which corresponds to a word that the text encoder already has in it's vocabulary. The training set is usually small (3 - 5 images), so this task can be considered as a *few-shot learning* approach. The implementation is a personally adapted version of the open source code for *Textual Inversion* that is publicly available in the *HuggingFace Diffusers Library*.

In the following sections, we study multiple aspects of this method, including ablation, qualitative, and quantitative studies and discussions. All these experiments were conducted using a single *GeForce RTX 2060 Super GPU* ($8GB$ of video memory), with a batch size of only 1 due to the limited amount of memory. Different values for the learning rate were tested prior to all the experiments, and the selected best value was 0.0005. Finally, it is of interest to mention that the version of *Stable Diffusion* that was used corresponds to a lighter one (uses half-precision floats), so this entire work can also be considered as a case study on image generation with low resources available, and considerably shorter training times.

4 Qualitative Ablation Study

In the following section, we perform an ablation study over the most relevant training parameters, discussing the results in a qualitative manner.

4.1 Number of Training Time-Steps

We begin by looking at the effect the amount of training time-steps has over the model performance on image synthesis. Here, each time-step refers to the optimizer algorithm updating the weights for the embedding of the new concept. This experiment is carried out for a total of 5000 steps, with the most interesting checkpoints looking as shown in **Figure 3**.

As our results demonstrate, the number of training time-steps has a clear effect on the generated images. In both examples, it takes at least 1000 steps to achieve a good level of reconstruction for the learned concept. Looking at the first example, the ideal number of steps seems to be around the 1500 mark, since after that, the concept starts showing up multiple times inside the images (in a distorted fashion), showing signs of overfitting. In the second example, it takes longer to adapt properly to the new style (around 1800 steps), and overfitting doesn't seem to happen (at least not in an intrusive way).

From these results, we can conclude that the number of training time-steps should be kept around the 1500 - 2400 range for optimal performance.

4.2 Initial Embeddings

Now, we analyze the effect of the initial embeddings used to learn the new concepts. We expect to obtain considerably different results when using different embeddings, due to the sensible nature of the low-dimensional latent space vectors (see **Figure 4** for results).

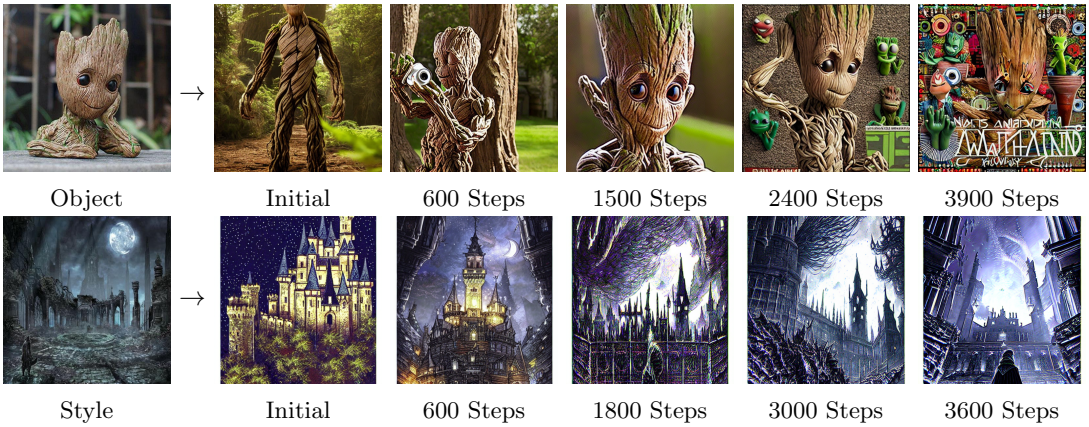

Figure 3: Effect of the number of training time-steps on generated images. In the first example, the concept is an **object**, corresponding to a small version of the fictional character *Groot*, from *Marvel*. In the second example, the concept is a **style**, corresponding to the video game *Bloodborne*.

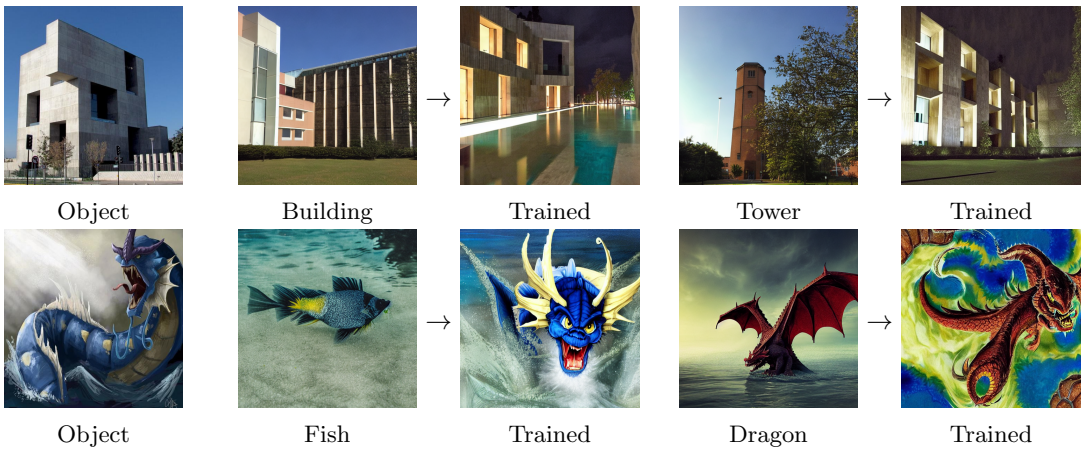

Figure 4: Effect of the initial embeddings used for training new concepts for objects. In the first example, the object corresponds to the *Innovation Center Building* from *PUC*, which is trained from the embeddings of the words **Building** and **Tower**. In the second example, the object is *Gyarados*, a fictional creature from the *Pokemon* franchise, which is trained from the embeddings of the words **Fish** and **Dragon**.

As seen in the results, our hypothesis was correct. In the first example, the difference may not be as noticeable, due to both initial embeddings (*Building, Tower*) being semantically similar, but the image obtained by training from the word *Building* shows a shorter building than the one in the image that comes from the word *Tower*, which makes sense, as a tower is a type of building that is supposed to be tall. On the other hand, for example 2, the difference is crystal clear. Using a *Fish* as the initial embedding produces a high quality image that soundly represents the desired creature (albeit in a different art style), while the initial embedding for *Dragon* ends up dominating over the object features, to the point where most of it's visual characteristics come from a dragon instead.

From these results, we can conclude that choosing the proper initial embeddings is key to obtaining the desired effect.

4.3 Randomization Seed

Another important parameter, as usual, is the randomization seed used for pseudo-random number generation throughout training.

In **Figure 5**, we can observe that using two different seeds can completely alter the final results. In the example, seed **A** delivers an image of a tree with an angry and unnatural face. In contrast,

Object Seed A Seed B

Figure 5: Effect of the randomization seed used for training a new concept. The object corresponds to the *Deku Tree*, a fictional character from the *Legend of Zelda* video games. The two generated images were obtained through training with different seeds.

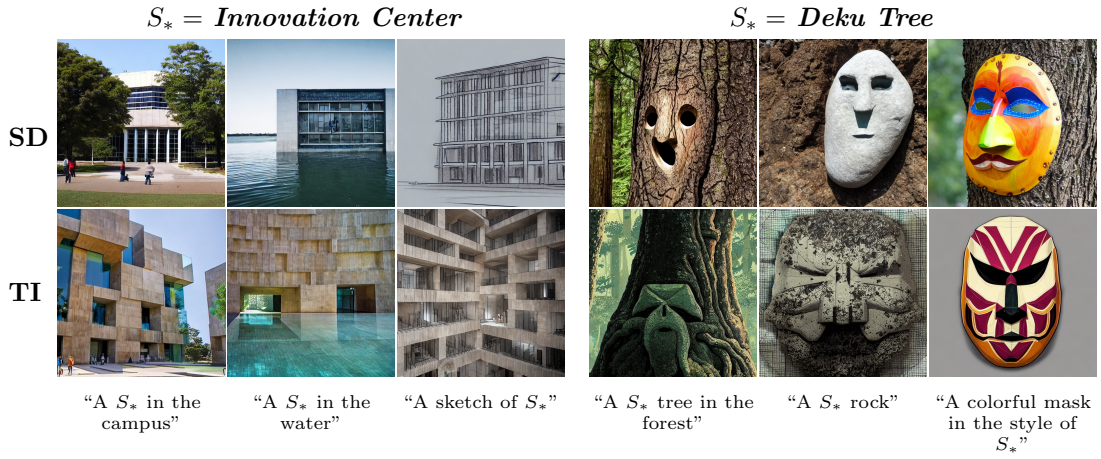

Figure 6: Comparison of expressive power between the *Stable Diffusion* baseline **(SD)** and *Textual Inversion* **(TI)**, on the task of **text-guided image synthesis**.

seed **B**, which also ends up with an image of a tree with a face, contains more realistic features that one could possibly find somewhere in the world (with a little luck).

From these results, we can conclude that trying different seeds is a good strategy to find better guidance in training, and it also allows us to obtain alternative generators for the same concept, adding more variety to our image synthesis arsenal.

5 Qualitative Analysis

One of the main practical applications of large scale text-to-image models is in the field of digital art composition. Accordingly, in this section we conduct a qualitative analysis over some of the more interesting samples obtained through experiments, to determine the potential *Textual Inversion* has in the task of producing art.

5.1 Text-Guided Synthesis

Text-guided synthesis refers to the task of composing novel scenes by incorporating the learned concepts from training into new conditioning textual prompts (for inference). To start, we'll compare the expressive power of the vanilla *Stable Diffusion* model **(baseline)** with that of the *Textual Inversion* (*TI*) approach. The idea is to prove that *TI* effectively improves over the baseline, due to the fact that it gives the user more fine-grained control over concepts present in the generated images (see **Figure 6**).

As the comparison shows, *TI* allows the user to introduce a specific concept of interest that would be impossible to properly describe in natural language for the baseline. In the first example (left), the vanilla model produces high quality images, but they contain a generic concept of a building. Meanwhile, *TI* generates contextually similar images to those of the baseline, but with the specific semantics of the desired type of building (*Innovation Center*). Similar observations can be taken

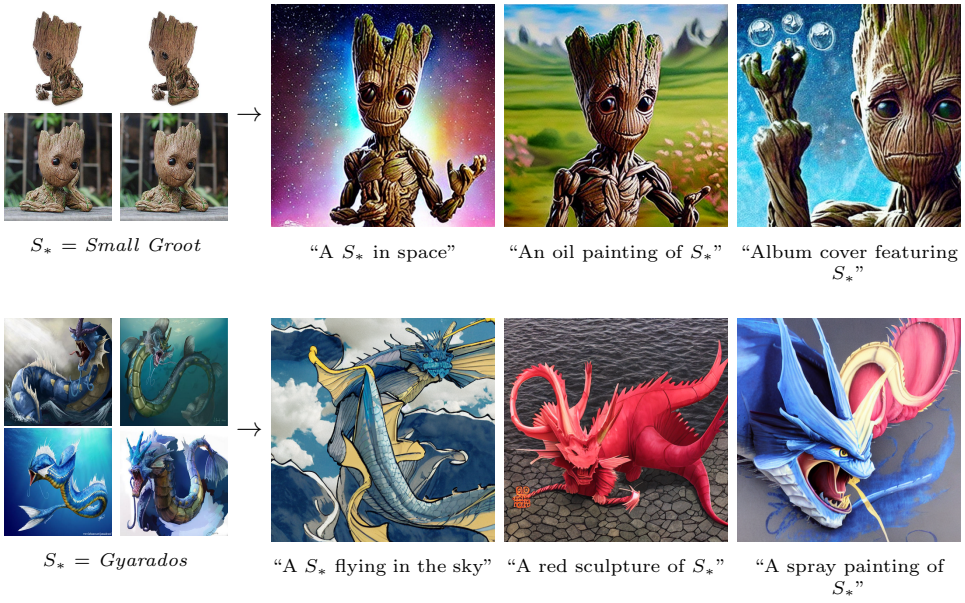

Figure 7: Examples of **text-guided image synthesis** for different novel objects using *Textual Inversion*, given a small dataset of examples.

from the second example, where the vanilla model fails to apply the correct face of the *Deku Tree* object, while *TI* learns it properly via lightweight *fine-tuning*. One thing of note here is the fact that the variety of the produced images seemingly decays when applying *TI*. This will be further discussed in **Section 6**.

So far, the results demonstrate that *TI* effectively augments the expressiveness and fine-grained control that the user commands through the conditioning textual prompts. With this, we can now take a look at the artistic capabilities of the approach.

The synthesized images from **Figure 7** showcase some of what the model is capable of. Tasks like placing a novel object in a different background, modulating the gestures of said object to fit the prompt, and mapping the object to a particular school of art are performed really well. Improvements could be achieved in the sense of adding more variety to what the object is doing, instead of closely reproducing the themes of the training set.

After discussing all these text-guided image synthesis results, we can safely conclude that *Textual Inversion* is well suited for this task, artistically speaking.

5.2 Style Transfer

Style transfer refers to the task of applying a specific art style to concepts that were not necessarily meant to be in said art style. This is a more abstract task than image synthesis, since the model has to learn to modulate details in the pictures while mostly maintaining the original semantics. Some interesting results are displayed in **Figure 8**.

As inferred from our results, the model successfully maps the desired concepts into the art style conditioned by the prompt. Due to the nature of this task, the generated images are actually even more rich in terms of variability than the ones from image synthesis. After briefly reviewing these style transfer results, we can conclude that *Textual Inversion* is also fit for this kind of more abstract tasks.

Finally, from the perspective of art quality, *Textual Inversion* is a really good candidate for automated image generation problems, since it can produce high quality results in both concrete and abstract tasks, with the only visible drawback being the observed loss of variety at specific examples.

Figure 8: Examples of **style transfer** for different styles and prompts using *Textual Inversion*, given a small dataset of examples.

6 Quantitative Analysis

The general notion obtained from the qualitative analysis, signals that one possible weakness of *Textual Inversion* could be the lack of variety due to fitting too much to the few-shot training data, or perhaps because of a learning path that omitted a substantial amount of relation-rich data from the embeddings, ending up with a model that has lost flexibility compared to the vanilla *Stable Diffusion* baseline. To evaluate this observation, we propose a brief quantitative study, based on the latent space representations from *CLIP* [**2**].

6.1 Evaluation Metrics

We consider two aspects of image generation quality: *reconstruction* and *editability*. Reconstruction refers to the model's ability to replicate the target concept from the training set, so in other words, we want a numerical score that represents image similarity between the generated images for a concept and the training set images of said concept. Editability refers to the model's ability to modify the semantics of the generated images using the textual prompts, which means that we need a score that measures similarity between conditioned images and the prompt that conditioned them.

In the case of reconstruction, the proposed score is image similarity between a set of 16 generated images and the images from the training set used for the concept of interest (could be any of the concepts we've explored in this work). This image similarity will be specifically measured with the average pair-wise cosine similarity between both sets of images, applied over the embeddings in *CLIP* latent space.

In the case of editability, we first generate three sets of 16 images each, where each set is conditioned on a specific prompt that tries to *edit* the novel concept. Afterwards, for one of these sets, we calculate the average embedding in *CLIP* latent space, and then measure the cosine similarity between that and the latent representation of the textual prompt associated with the current set of images. This step is applied for all three sets and their respective conditioning prompts, with the proposed score as the average of the three resulting values.

6.2 Results

The proposed metrics were evaluated for both the vanilla *Stable Diffusion* model and the *Textual Inversion* model. Results are summarized in **Table 1**.

Model	Reconstruction	Editability
SD	0.5622	0.2639
SD + TI	0.8077	0.1611

Table 1: Results for metric evaluation of model capabilities.

As expected, the score for reconstruction increases considerably when using *TI*, which is supported by the qualitative study comparisons. On the other hand, editability does decrease when applying *TI*, which confirms the suspicions that were discovered before. In summary, the model clearly presents a trade-off between both metrics, though the increase in reconstruction capacity is larger and thus *TI* is definitely justified as a powerful tool for custom image generation, as stated previously in multiple occasions.

7 Conclusions

In this work, we studied, tested and analyzed different aspects of a method known as *Textual Inversion*, which leverages a large scale pre-trained text-to-image model to create images of specific concepts in novel settings and scenes. The method is simple in it's execution, and provides the user with an interface that allows them to input visual cues into an image generation model, so that the limitations of natural language input prompts can be avoided when dealing with novel concepts that the baseline approach doesn't handle well.

For the purposes of this work, the selected baseline was *Stable Diffusion*, a model that belongs to the category of *LDM* approaches. Using this baseline, a multitude of experiments were conducted, including an ablation study, a qualitative analysis in terms of artistic potential, as well as a brief quantitative experiment to further discuss possible advantages and weaknesses of the *Textual Inversion* strategy. From the obtained results, we conclude that this approach is in fact, an excellent choice for tasks such as text-guided image synthesis and style transferring.

In terms of future work, there is plenty to investigate. For example, the effect of regularization techniques, higher capacity models, and high-performance hardware for training can potentially remedy the problem of overfitting, as well as help increase editability. Dataset characteristics could also be studied, to try defining a standard type of dataset that helps with convergence towards the desired concept when training. Finally, there are more tasks that need to be tested and enhanced, specifically, ones that require precise logical or compositional knowledge, since the current approach would fail to learn this with just a couple images of training. More over, compositional relations between multiple concepts in the same image would be an interesting objective of study, maybe requiring the use of novel types of training sets for this tasks.

References

[1] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. 2022.

[2] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sustskever. Learning transferable visual models from natural language supervision. 2021.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. 2021.

[4] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.