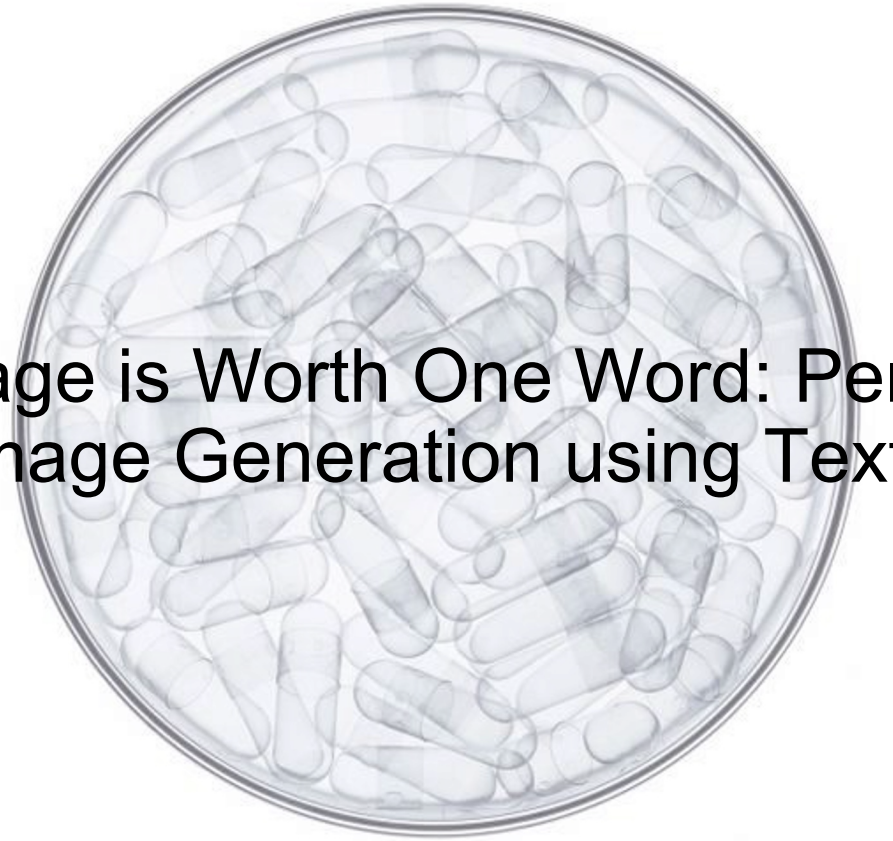


An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Agenda

- Introduction
- Background
- Motivation
- Method
- Evaluation
- Summary and future direction

Introduction

- In a famous scene from the motion picture “Titanic”, Rose makes a request to Jack:

  “...Draw me like one of your French girls”.

- Text to image generation tools have been used for artistic creation, as sources of inspiration, and even to design new, physical products .

- Text-to-image models allow users to synthesize novel scenes with unseen compositions and produce vivid pictures in a myriad of styles.
- 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.
- 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 propose to overcome these challenges by finding **new words in the textual embedding** space of pre-trained text-to-image models.
- 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.

Input samples $\xrightarrow{invert}$ "S_*"

"An oil painting of S_*"

"App icon of S_*"

"Elmo sitting in the same pose as S_*"

"Crochet S_*"

Input samples $\xrightarrow{invert}$ "S_*"

"Painting of two S_* fishing on a boat"

"A S_* backpack"

"Banksy art of S_*"

"A S_* themed lunchbox"

Methodology

We have various architectures followed for text to image model co having their own limitations.

The Current Architecture consists of the following three parts

- Latent Diffusion Model
- Text embedding
- Textual inversion

What is Latent Diffusion Model?

Diffusion models work by replicating this diffusion process by adding noise to original images and later learning how to reverse this noise process following the same steps in reverse manner. the noise here is applied to the images following a markov chain.

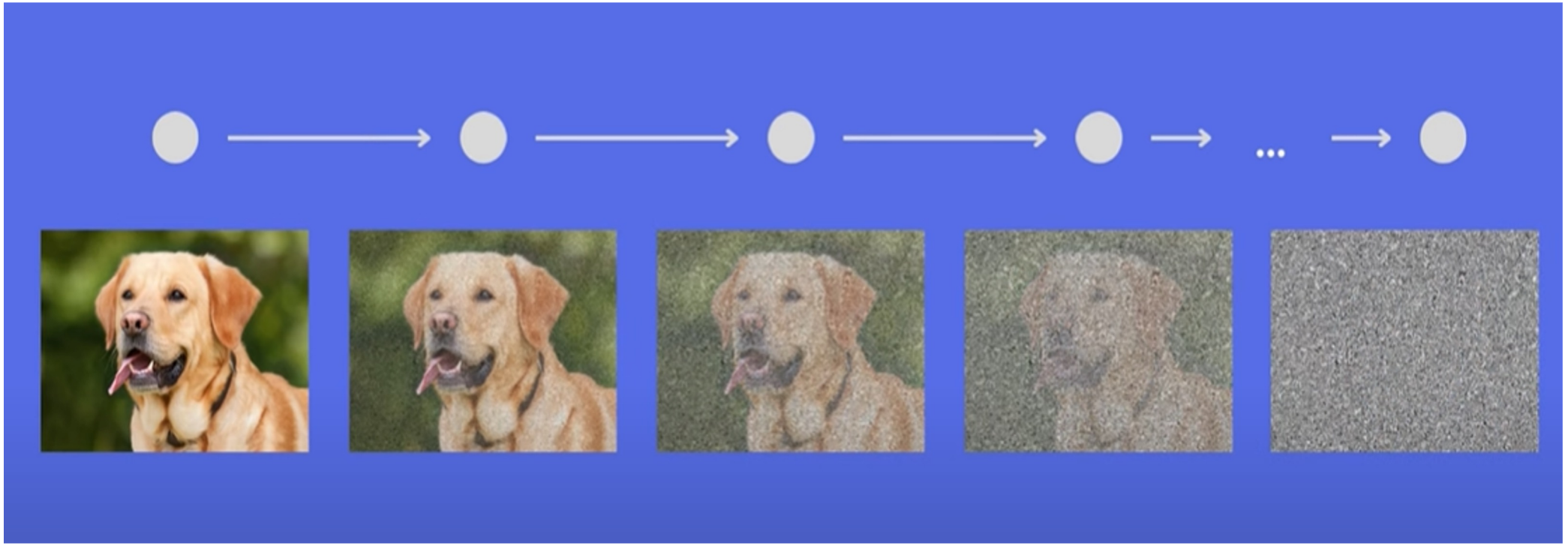
What is a markov chain?

A markov chain is a chain of events where the current time step only depends on the previous time step so that means there are no cross dependencies between time steps that do not immediately follow each other.

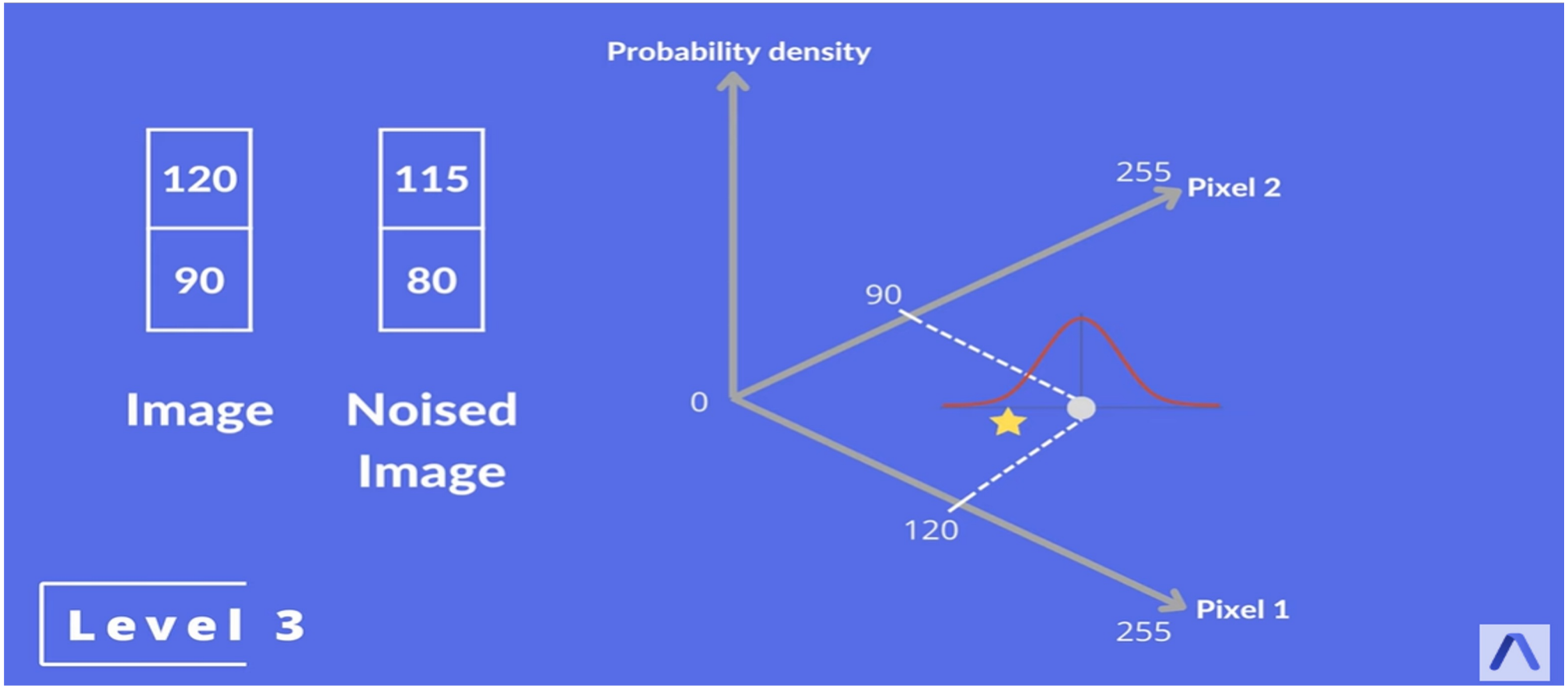
What is Latent Diffusion Model?

The procedure that involved in markov chains make it tractable for the noise adding to be reversed later so at the end a diffusion model is a Markov chain where in each time step we add a little bit of noise to our image until the image only consists of noise and later learning how to reverse this noise adding process after it is trained given only noise this model is able to generate high resolution images level.

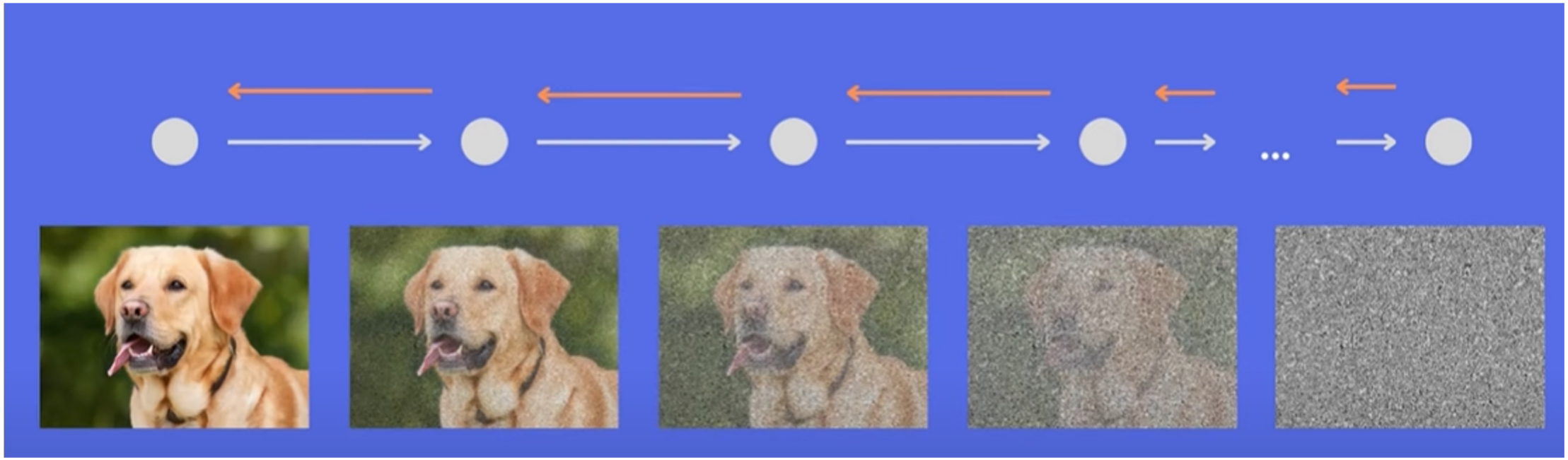
What is Latent Diffusion Model?



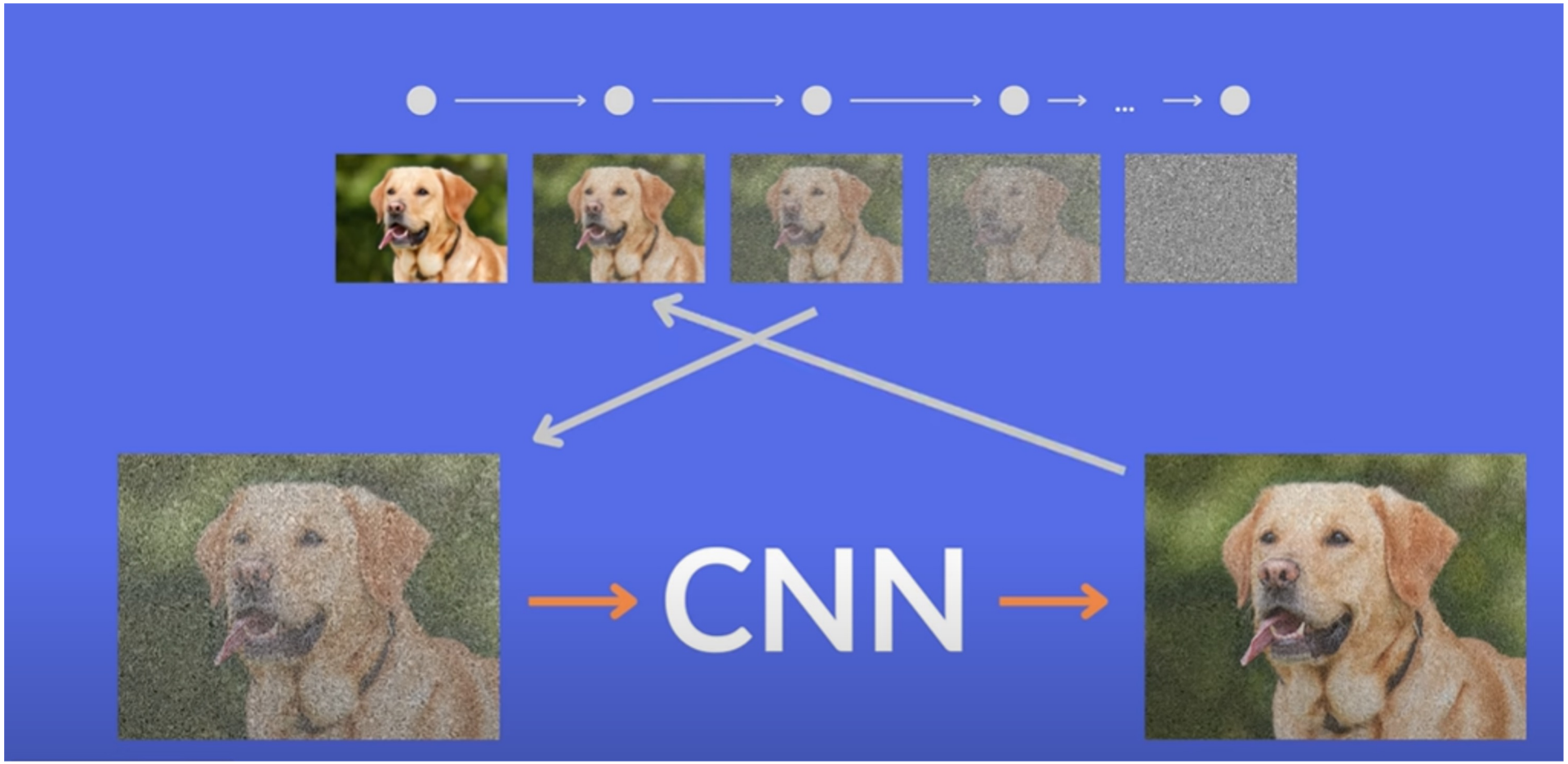
What is Latent Diffusion Model?



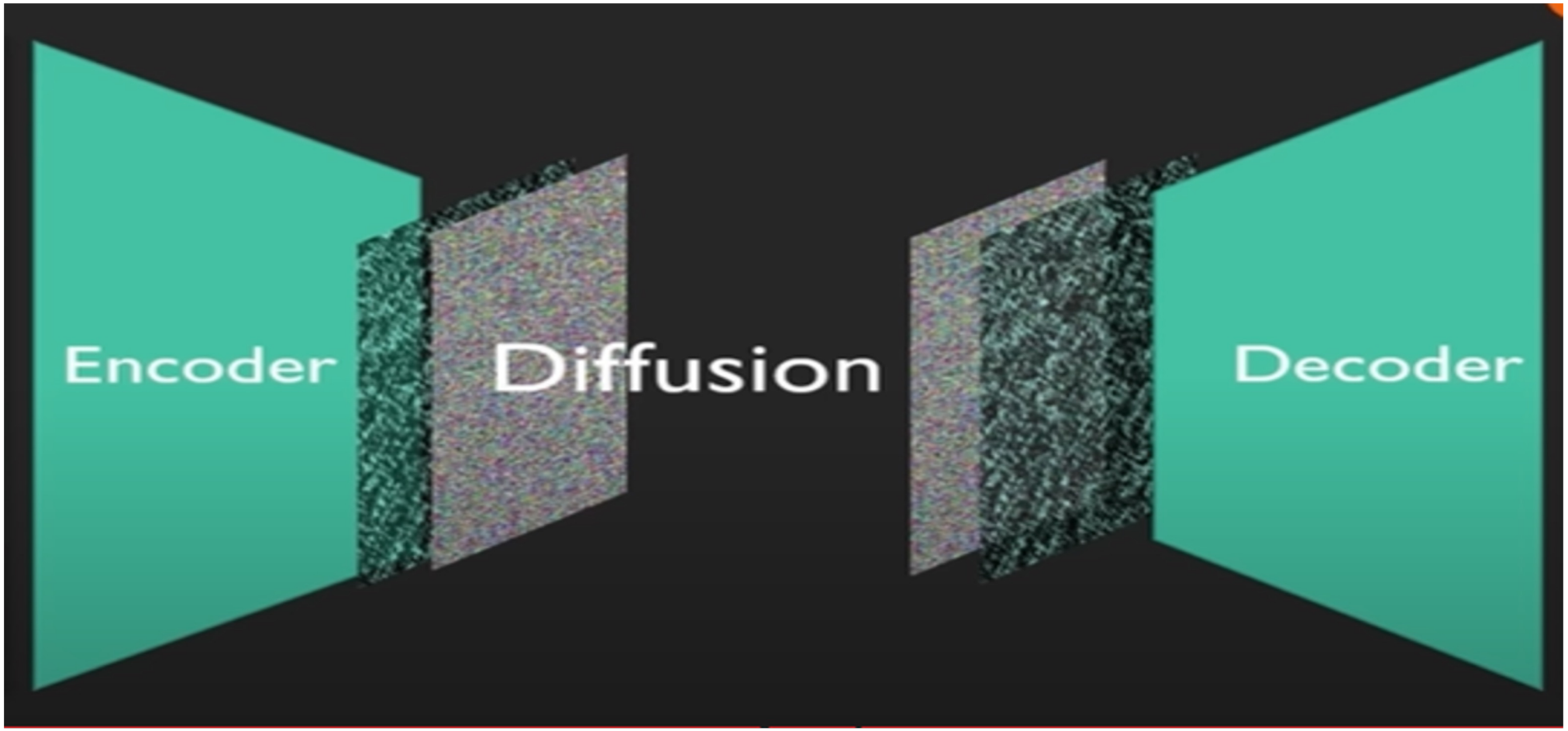
What is Latent Diffusion Model?



What is Latent Diffusion Model?



What is Latent Diffusion Model?



What is Text Embedding ?

Text embedding is a technique in machine learning and natural language processing (NLP) that represents words or phrases in a numerical format so that they can be processed by machine learning models. In text embedding, each word or phrase is mapped to a high-dimensional vector that captures its meaning, context, and relationships with other words in the text.

What is Textual Inversion ?

Textual Inversion is a technique for capturing novel concepts from a small number of example images in a way that can later be used to control text-to-image pipelines. It does so by learning new ‘words’ in the embedding space of the pipeline’s text encoder. These special words can then be used within text prompts to achieve very fine-grained control of the resulting images.

Implementation

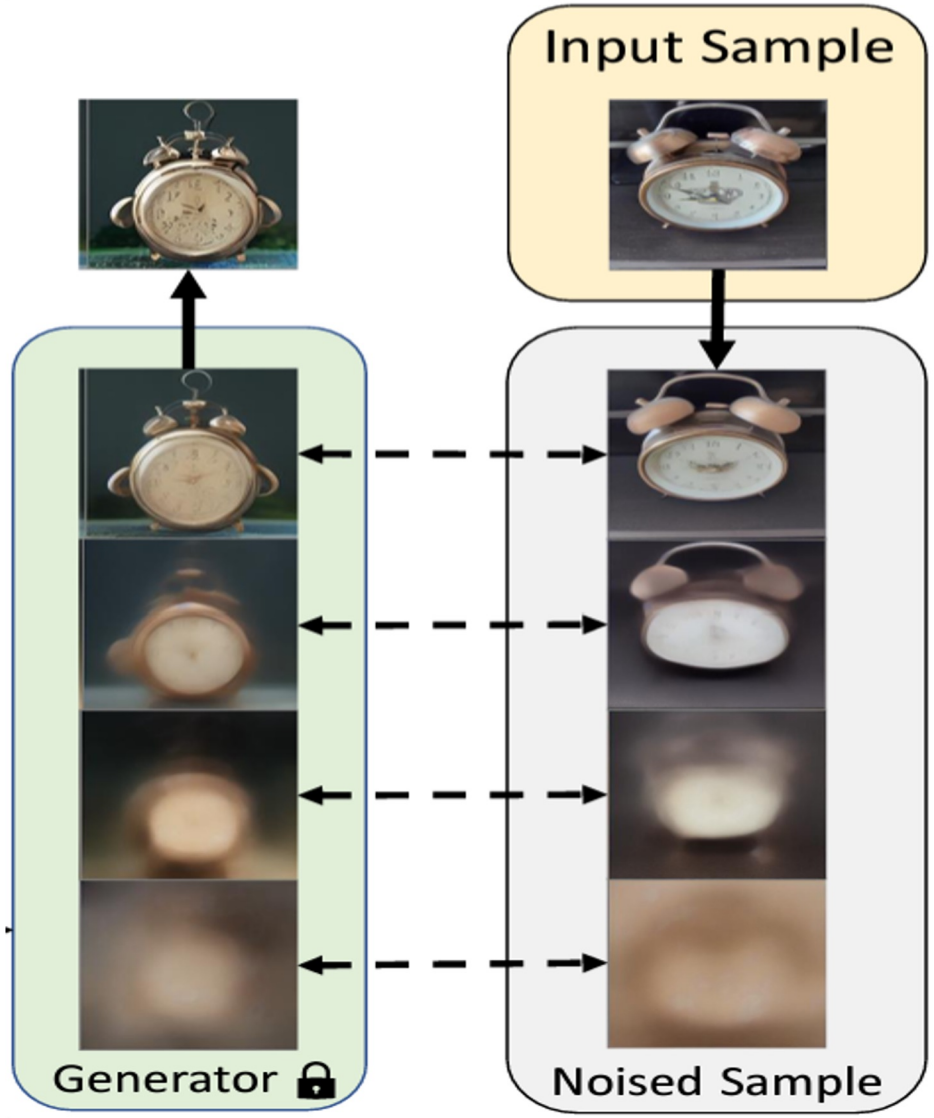
Consider we have a text prompt can be used in a diffusion model, it must first be processed and go through the following steps

- Typical text encoder models, such as BERT, start with a text processing step.
- 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.

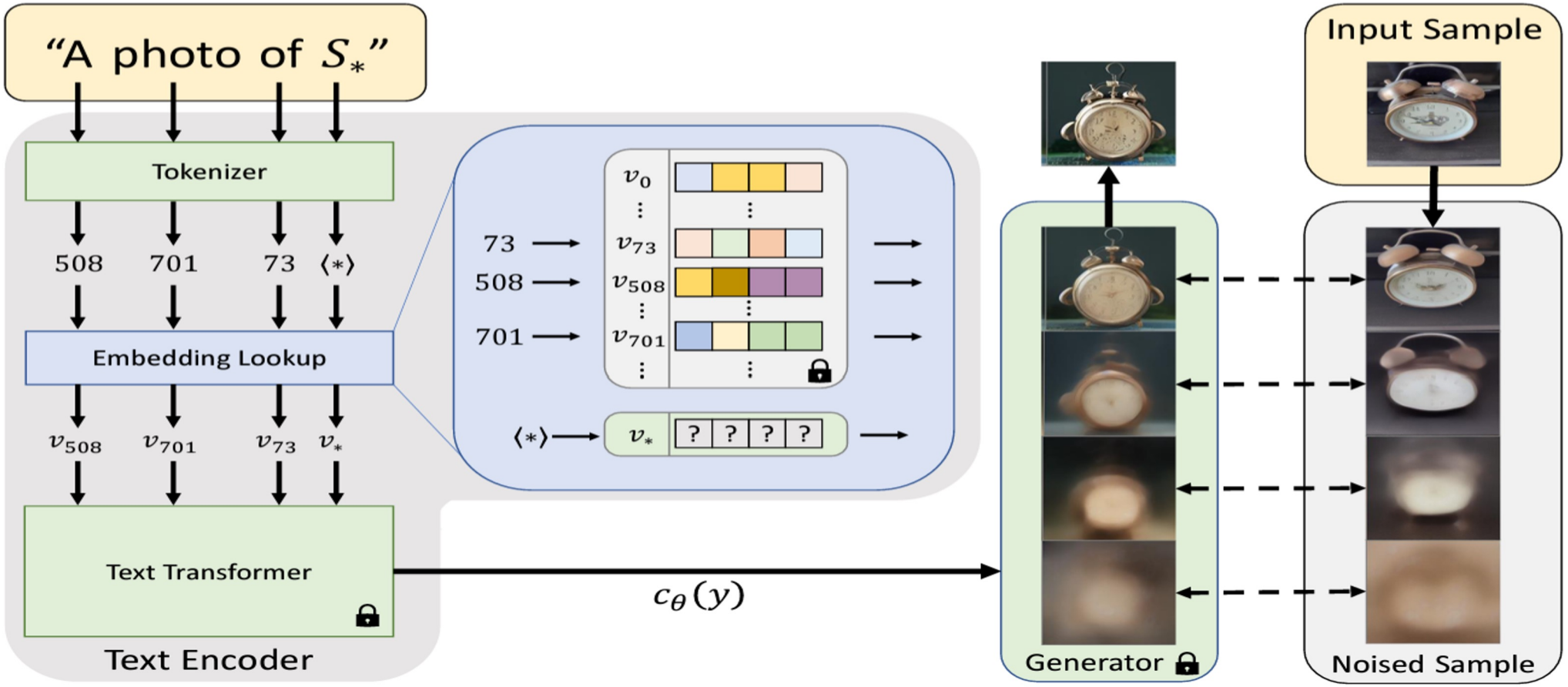
Implementation

- 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.
- And Textual inversion 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.
- It is then used in conjunction with a noised version of one or more training images as inputs to the generator model, which attempts to predict the denoised version of the image.

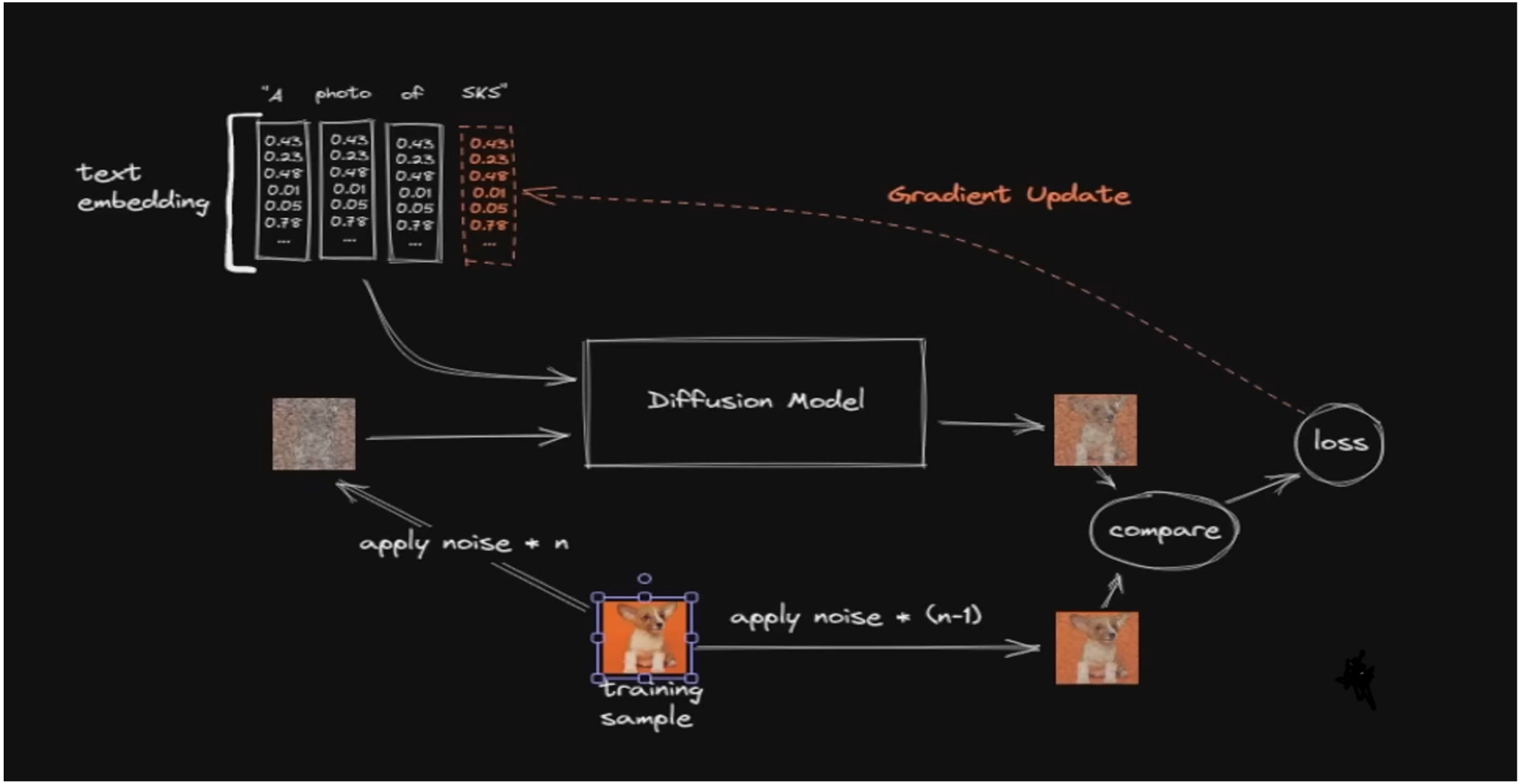
Implementation



Implementation



Implementation



Example

Input samples

"S_* sports car"

"S_* made of lego"

"S_* onesie"

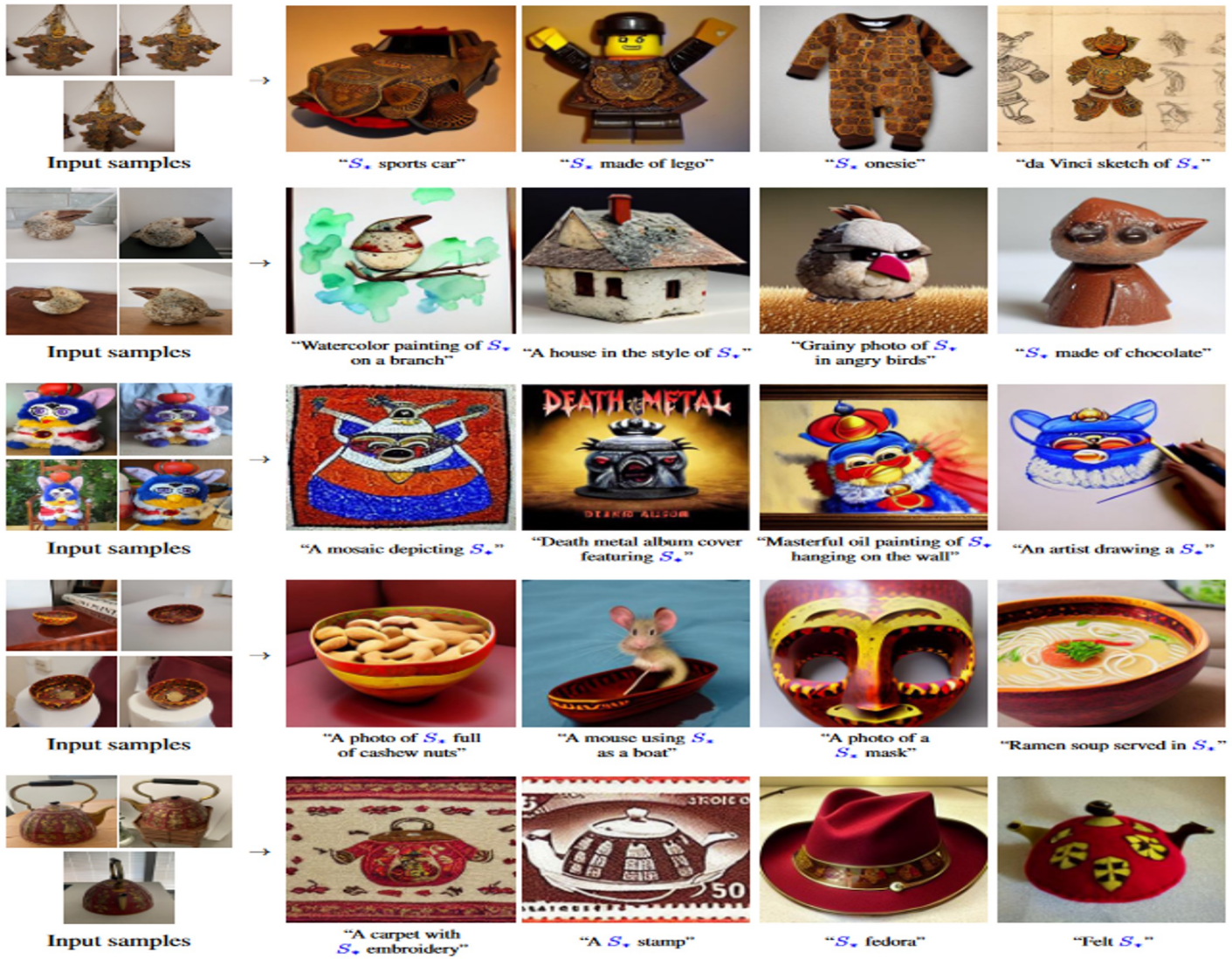
"da Vinci sketch of S_*"

Qualitative comparisons and applications

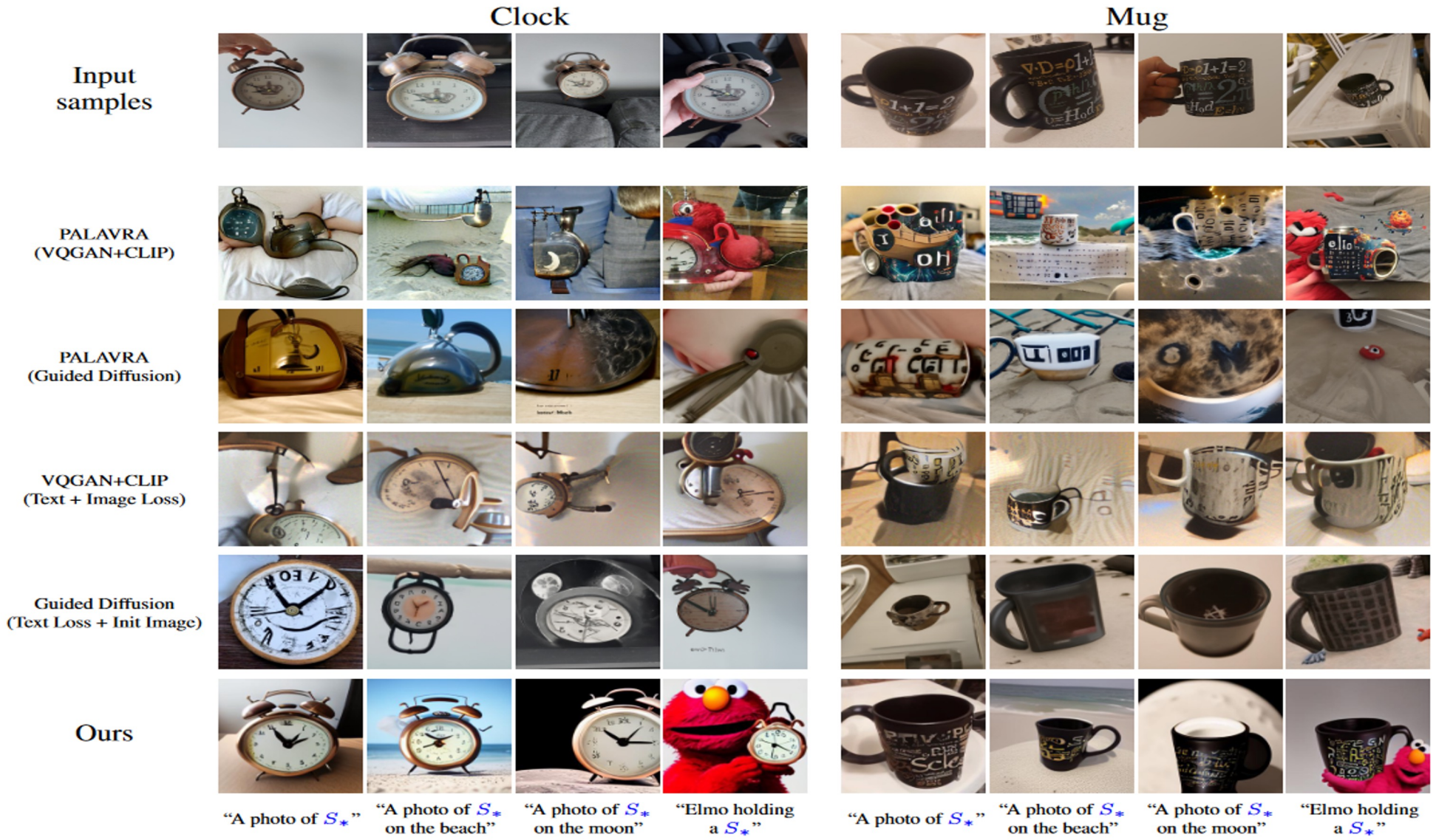
1. Image variations

Our method generates variations which are typically more faithful to the original subject

2. Text-guided synthesis



Additional text-guided personalized generation results.



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.

3. Style transfer

The textual-embedding space can represent more abstract concepts, including styles.

4. Concept compositions

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.

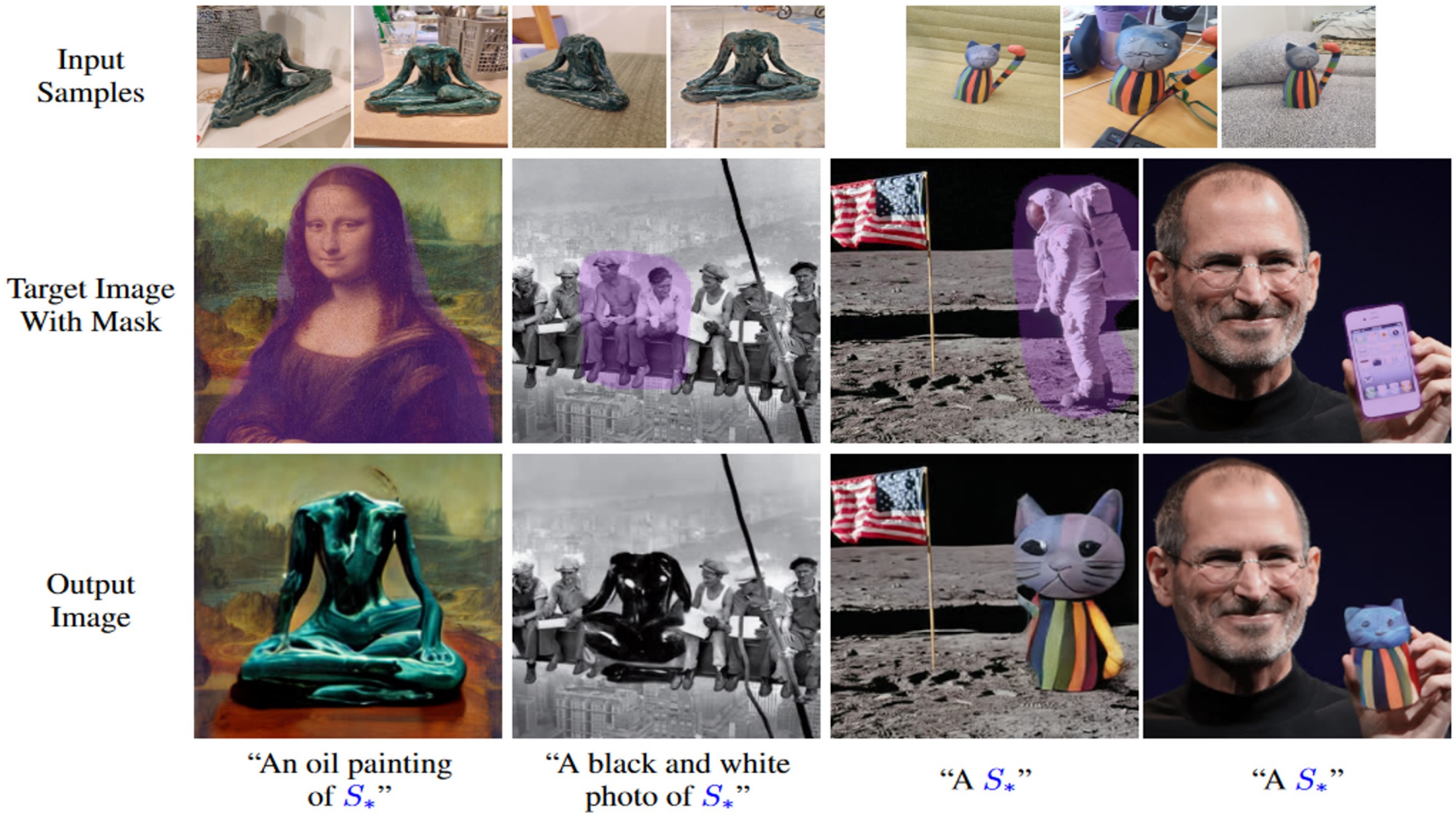
5. Bias Reduction

“A stock photo of a doctor” (Base model)

“A photo of S_*” (Ours)

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.

6.Downstream application

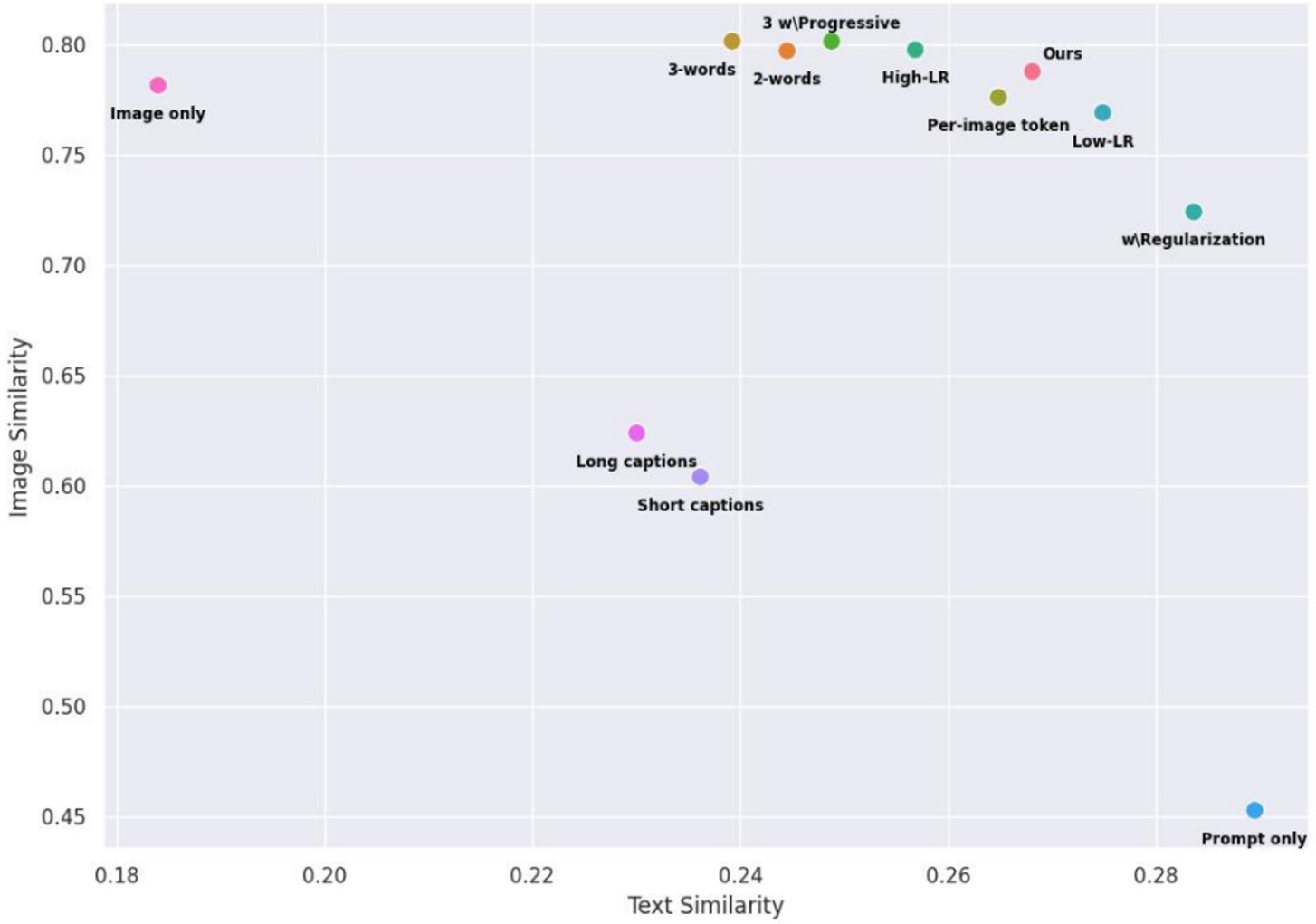


Our words can be used with downstream models that build on LDM.

Quantitative Analysis

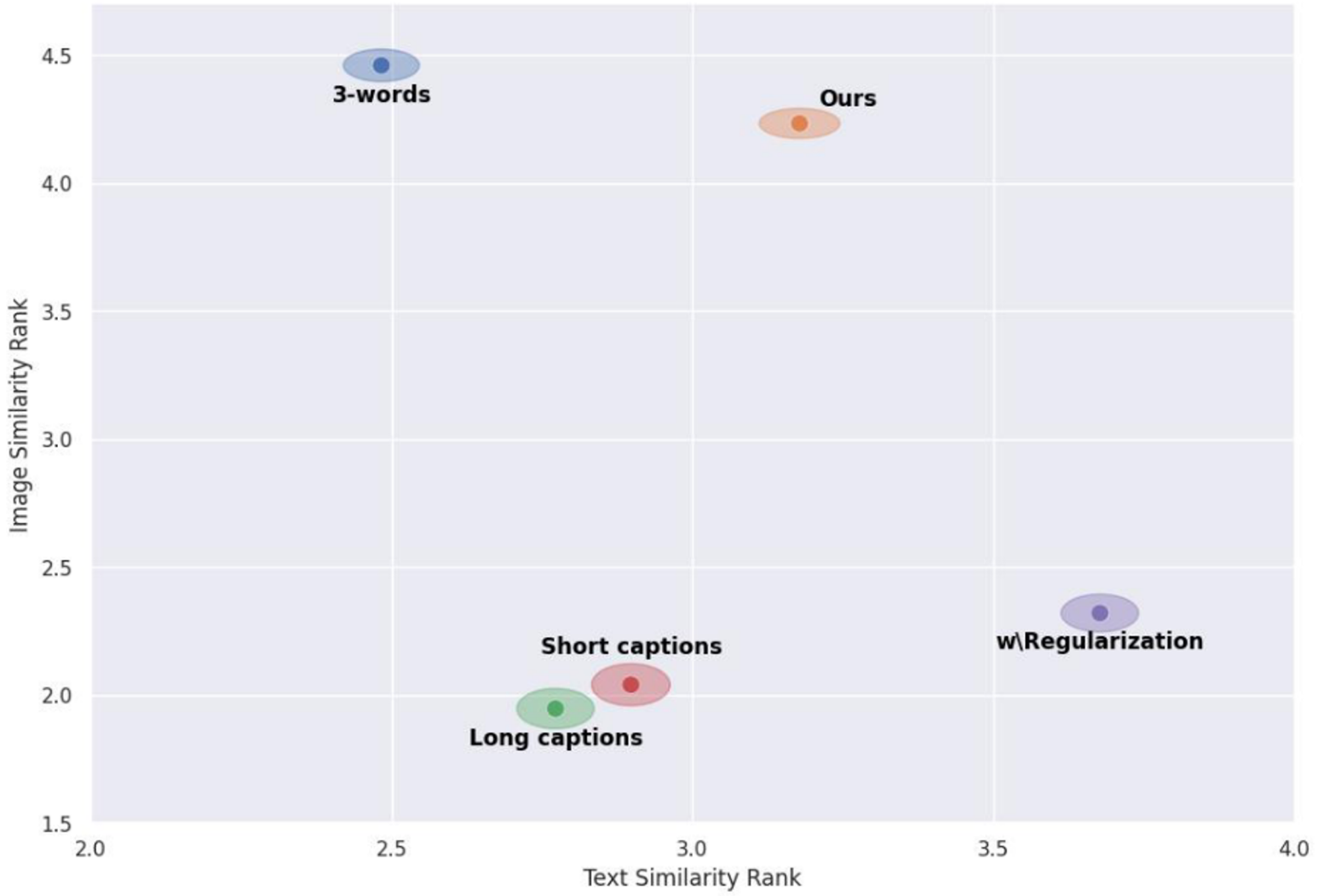
- **Extended latent spaces :** we consider an extended, multi-vector latent space. In this space, $S*$ is embedded into multiple learned embeddings, an approach that is equivalent to describing the concept through multiple learned pseudo-words.
- **Progressive extensions** :Here, we begin training with a single embedding vector, introduce a second vector following 2, 000 training steps, and a third vector after 4,000 steps.
- **Regularization:** 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 .
- **Per-image tokens :**We investigate a novel scheme where we introduce unique, per-image tokens into our inversion approach.
- **Human captions :** We compare to human-level performance using the captions.
- **Textual-Inversion :**we consider our own setup. We further evaluate our model with an increased learning rate.

Results


3 w\Progressive
3-words
2-words
High-LR
Ours
Image only
Per-image token
Low-LR
w\Regularization
Long captions
Short captions
Prompt only
Image Similarity
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.18
0.20
0.22
0.24
0.26
0.28
Text Similarity

(a)

Results


3-words
Ours
Short captions
Long captions
w\Regularization
Image Similarity Rank
Text Similarity Rank
4.5
4.0
3.5
3.0
2.5
2.0
1.5
2.0
2.5
3.0
3.5
4.0

Limitations

- While our method offers increased freedom, it may still struggle with learning precise shapes, instead incorporating the “semantic” essence of a concept.
- Lengthy optimization times is drawback . Single learning concept requires two hours.

Future Work

- Improving Image Quality: Further research is needed to improve the quality of the generated images and to increase their similarity to the textual descriptions. This could involve incorporating additional information, such as attributes or scene context, into the model.
- Incorporating Personalization: Research can be done to explore more sophisticated and comprehensive ways of incorporating personalization into the images, beyond just modifying the textual descriptions.

Future Work

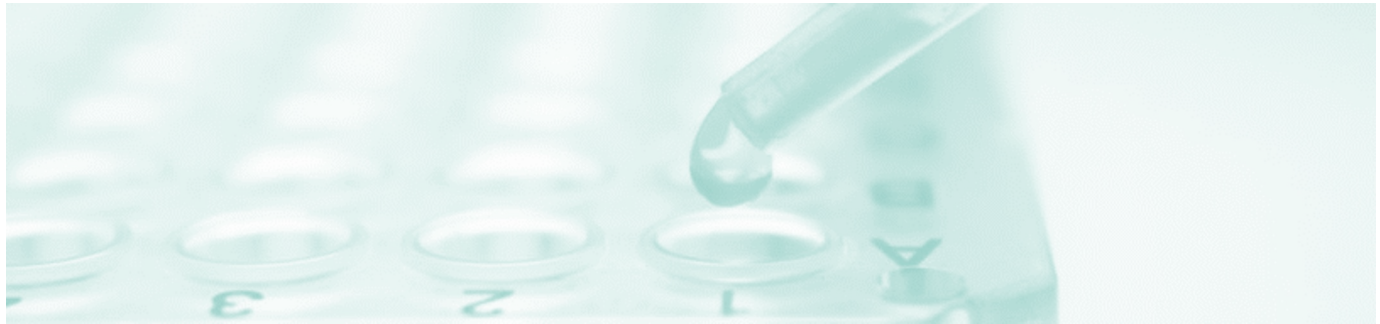
- Enhancing Interactivity: The proposed solution can be extended to allow for more interactive and user-driven text-to-image generation. This could involve incorporating user feedback or allowing users to directly manipulate the generated images.
- Exploring Different Architectures: Alternative model architectures, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), or Transformer-based models, can be explored to see if they lead to improved performance or better results.

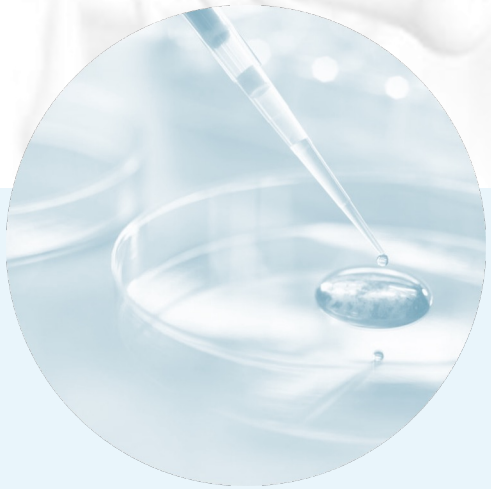
SUMMARY

The task of language guided image generation is introduced where text to image model is leveraged to create an image of specific concept in novel scene.

“Textual Inversions”, operates by inverting the concepts into new pseudo-words within the textual embedding space of a pre-trained text-to-image model is implemented in LDM.

We hope this approach paves the way for future personalized generation works.





THANK YOU