Personalized Meme Image Generation via Text Inversion

15.778 - Final Report

Kaibo Zhang	Annie Liu	Lauren Zhang	Tong Lyu
kbzh2558@mit.edu	yil297@mit.edu	lzhang27@mit.edu	tonglyu@mit.edu

December 10, 2025

1 Introduction

Large text-to-image diffusion models such as Stable Diffusion have demonstrated remarkable generative capabilities, yet they still struggle to represent user-specific concepts that fall outside the scope of their pretrained vocabularies. Direct fine-tuning of these large models is computationally prohibitive and raises privacy concerns in consumer-facing settings.

Textual inversion [1] offers an efficient alternative by introducing a pseudo-token that is optimized solely through its embedding, allowing new visual concepts to be learned while keeping all diffusion model parameters frozen. In this work[1] , we applied textual inversion to Stable Diffusion v1.5 to enable the creation of personalized memes and reaction figures from only a few user-provided images. We studied both the classical single-vector formulation and an extended multi-vector variant that expands the embedding capacity and aims to capture more nuanced stylistic and structural cues. Our findings highlight the limitations of the one-vector representation and demonstrate that a multi-vector approach yields substantially

[1]Project repository available at:
`https://github.com/kbzh2558/memes_generation_via_text_inversion`
Collab Notebook for single vector available at:
`https://colab.research.google.com/drive/1i7Z1A4IQnbJl0xoyY0YlaFTvB0RmQ9Fa?usp=sharing`
Collab Notebook for multi-vector available at:
`https://colab.research.google.com/drive/1uqVRKLv8UQVk3ty3i551SUxuzfCrZAYo?usp=sharing`

stronger generalization, improved identity retention, and greater robustness under prompt variation.

2 Data Description

Our dataset construction was motivated by a recent viral streaming meme that circulated widely across major social media platforms in mainland China, featuring a stylized reinterpretation of the Madagascar penguin character dressed in formal attire. This meme provided a natural test case for examining whether textual inversion can capture a culturally emergent visual concept from only a few examples. We curated four representative images of the same character, chosen to preserve semantic identity while varying pose, viewpoint, and the objects held (Figure 1). This small but diverse collection was intended to encourage the model to extract stable conceptual features rather than memorize individual instances.

All images were checked for stylistic consistency and standardized to 512×512 pixels. Preprocessing followed established practices for concept learning in diffusion models, including resizing, center cropping, and mild color jittering to reduce sensitivity to local pixel statistics. These design choices yield a compact yet informative dataset that supports a controlled evaluation of how effectively single vector and multi vector textual

3 Methodology

3.1 GD in Text Inversion and Stable Diffusion

In order to replicate the textual inversion procedure of Gal et al. (2022), we optimized a single learnable embedding vector associated with a newly introduced placeholder token within the Stable Diffusion v1.5 text encoder, while keeping the VAE, UNet, and base text encoder weights frozen. Given four concept images, we constructed a prompt-diversified training set by randomly sampling caption templates of the form "a photo of `<style>`" or syntactic variants thereof, ensuring that the placeholder token appears in multiple linguistic contexts as prescribed in the original paper (Table 1). Following empirical findings, we also initialized the learnable token using a word (i.e., "penguin" in our case) whose embedding best reflects the object depicted in our training images, which improves convergence and stabilizes early training dynamics. The embedding was trained for a total of 3000 optimization steps using

the Adam optimizer with a learning rate of 5×10^{-4} and a cosine warmup schedule over the first 10% of steps. Each iteration encoded an input image x into a latent z, perturbed it with noise ϵ at a randomly selected diffusion timestep t, and passed it through the UNet conditioned on text encoder outputs containing the learnable embedding. The embedding vector $\mathbf{v}$ was optimized on a Colab T4 GPU by minimizing a diffusion noise-prediction objective composed of an MSE reconstruction term together with an L_2 norm-regularization penalty of strength $\lambda = 10^{-3}$ to preserve the magnitude of the initializer embedding. The resulting optimization problem is

$$\mathcal{L}(\mathbf{v}) = \|\hat{\epsilon}_\theta(z_t, t, \mathbf{v}) - \epsilon\|_2^2 + \lambda \left(\|\mathbf{v}\|_2 - \|\mathbf{v}_0\|_2\right)^2,$$

where $\mathbf{v}_0$ is the original initializer embedding. Only $\mathbf{v}$ receives gradients, allowing the model to encode the target concept consistently across prompt variations.

3.2 Extension to Multi-Vector Representation

Recognizing a key drawback of Stable Diffusion, namely that its powerful text–image prior often dominates the influence of a single learned token, we observed that single-vector textual inversion struggled to sufficiently override these priors, resulting in suboptimal reconstructions of our target concept. We therefore extended the approach to a multi-vector representation, in which the concept is encoded not by a single embedding vector $\mathbf{v} \in \mathbb{R}^D$ but by a set of K trainable vectors

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1^\top \\ \mathbf{v}_2^\top \\ \vdots \\ \mathbf{v}_K^\top \end{bmatrix} \in \mathbb{R}^{K \times D},$$

each associated with a distinct placeholder token. The full placeholder phrase expands to "$*1\ *2\ \cdots\ *K$," allowing the concept to distribute across multiple embedding directions and thereby overcome the rigidity of the model prior. This construction can be interpreted analogously to multi-head attention, where each vector $\mathbf{v}_k$ captures a different semantic subcomponent of the concept, enabling the model to decompose appearance, texture, shape, and contextual cues across multiple embedding "heads." As in the single-vector case, we

initialized all vectors using the embedding of an initializer word chosen to best represent the object in the images, and added a small perturbation to encourage specialization. Training proceeded with 4000 steps and same hyperparameters as before, and we used exactly the same prompt templates described in Table 1. During the text-encoder forward pass, each placeholder token was replaced by its corresponding learned vector, and all vectors jointly optimized the diffusion noise-prediction loss with norm regularization. This multi-vector formulation mitigates the strong-model-prior issue and enables the learned concept to be represented with greater flexibility in the embedding space.

4 Results

4.1 Overfitting Risks of Unary Training Prompt

We initially trained the inversion embedding using a single fixed prompt of the form "a photo of `<style>`". However, we find that relying on a single fixed prompt causes the learned embedding to overfit to this exact phrasing. When the model only ever sees one phrasing during training, it learns to associate the concept token exclusively with that exact textual pattern. As a result, the embedding does not generalize to natural variations of phrasing. For example, prompts like "a picture of `<style>`" or "a close-up photo of `<style>`" produce weak or incorrect generations. To address this, at every training iteration we sample a caption uniformly at random from a small set of prompt templates (Table 1). These templates introduce mild but meaningful linguistic variation (e.g., "a full body photo of `<style>`", "a cropped photo of `<style>`", etc.). Introducing this mild diversity during training encourages the model to associate the learned concept with the visual content rather than memorizing the exact wording, resulting in more robust and prompt-flexible embeddings.

4.2 Superior Performance of the Multi-Vector

We trained both a single vector and a multi vector inversion embedding, with the 4-vector configuration chosen to enable each vector to specialize in different aspects of the concept such as pose, contour, texture, or accessory details. As shown in Table 2, this multi-vector formulation produces outputs that are consistently more stable and more faithful to the intended visual identity than the single vector baseline. Many generations produced by the

single vector embedding fail to reproduce the characteristic three-dimensional form of the sunglasses and often flatten the mouth and nose geometry, a phenomenon that likely reflects the difficulty of encoding multiple tightly coupled spatial and shading cues into a single direction in embedding space, whereas the multi-vector model is able to disentangle these factors and preserve their full 3D expression. As shown in the stuffed-toy example, where it preserves the character's form and key identity features far more reliably.

A clear improvement appears in prompts that require structural understanding or large changes in appearance. For example, the pencil-sketch prompt highlights how the 1-vector model can drift and produce unrelated scenes, while the 4-vector model keeps the correct character shape and pose. Together, these results suggest that the additional embedding capacity helps the model store richer concept information, leading to stronger identity retention and more consistent outputs across different styles. This improvement is plausible because a single vector must collapse all geometric, textural, and stylistic cues into one direction in embedding space, which limits its ability to encode interactions between shading, depth, and local structure. By contrast, the multi vector embedding distributes these factors across multiple learned directions, allowing the model to represent subtler variations in form and appearance that are essential for maintaining structural fidelity under style shifts.

4.3 Signal Loss with Reduced Representation

We tested how well the 4-vector representation handles long and descriptive prompts. Table 3 shows several captions that follow the same general theme, an oil painting of `<style>` in Paris, but each caption adds more details and becomes longer. The shortest prompt has about 12 words, while the longest includes more than 160 words.

With short or medium-length prompts, the 4-vector model works well. The `<style>` character appears clearly and stays consistent across samples. However, as the prompt becomes longer and adds many extra scene details, we notice that the `<style>` identity begins to weaken. The model shifts more attention to background descriptions such as the Eiffel Tower, lighting, city streets, or artistic style cues. In the longest prompts, the character may appear less detailed. This shows that even with four vectors, the learned concept can still lose influence when the prompt contains too many competing words. Long prompts

spread the model's attention across many tokens, and the `<style>` representation can be overshadowed.

5 Discussion and Conclusion

Our experiments show that Textual Inversion provides an efficient and practical mechanism for personalizing large diffusion models while avoiding the computational overhead of full-model fine-tuning. Across evaluations, the multi-vector formulation consistently outperformed the single-vector baseline, offering improved structural fidelity, better preservation of 3D geometry, and higher resilience to stylistic shifts. In addition to recovering fine-grained identity cues such as facial depth, accessories, and contour structure, the multi-vector embedding also demonstrated stronger stability under domain transfer tasks, including pencil sketches, stuffed-toy reinterpretations, and flat-cartoon renderings. Despite these advantages, our study also revealed a systematic failure mode: as prompts grow longer or include many competing scene descriptors, the influence of the learned concept weakens and can be overshadowed by background elements. This suggests that localized embeddings remain sensitive to the attention distribution of the underlying diffusion model and may struggle when the prompt implicitly reallocates semantic weight away from the concept token. Future work may address these limitations by exploring cross-attention modulation, adaptive prompt weighting, or hybrid strategies that combine textual inversion with lightweight parameter-efficient fine-tuning to improve identity retention under complex, multi-element prompts.

A Graphs and Tables

Fig. 1

Fig. 2

Fig. 3

Fig. 4

Figure 1: Illustration of Training Data

ID	Prompt Template
1	`a photo of {}`
2	`a full body photo of {}`
3	`a close-up photo of {}`
4	`a cropped photo of {}`
5	`a portrait of {}`

Table 1: Prompt templates used for object-based textual inversion training.

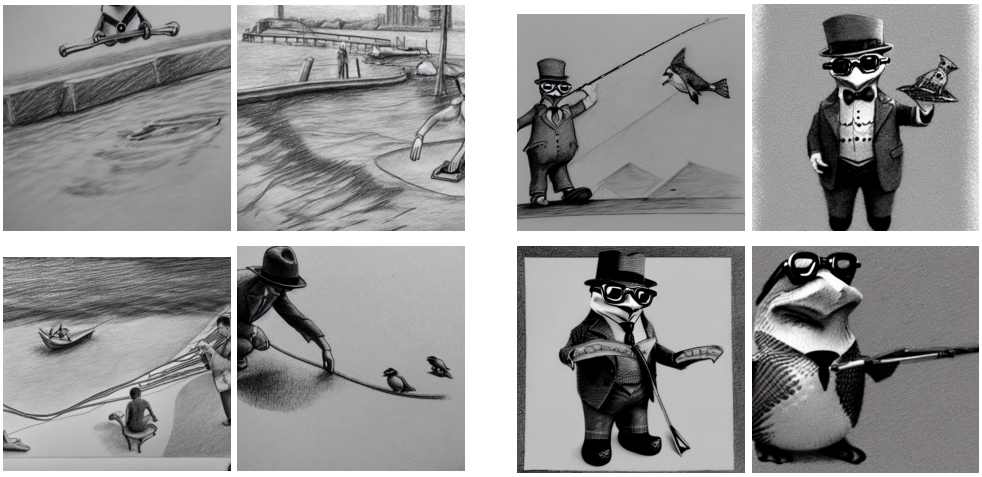
Prompt	Single Vector	Multi-Vector
A stuffed toy of `<style>`		
A pencil sketch of `<style>` fishing		
Watercolor painting of `<style>`		

Table 2: Comparison of outputs for different text prompts

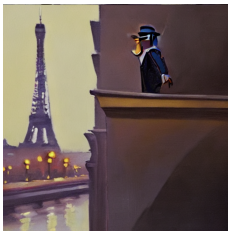
Prompt	Generation 1	Generation 2
An oil painting of `<style>` in Paris looking at the Eiffel Tower. [12 words]		
An oil painting of `<style>` in Paris looking at the Eiffel Tower at dusk, with warm colors and soft brushstrokes. [20 words]		
An oil painting of `<style>` in Paris looking at the Eiffel Tower during a vibrant evening, surrounded by city lights . . . impressionist textures. [26 words]		
An oil painting depicting `<style>` in Paris looking at the Eiffel Tower . . . through a complex interplay of style and composition. [50 words]		
An oil painting depicting `<style>` in Paris looking at the Eiffel Tower . . . shifts at the horizon—serve to enrich the composition without altering the underlying stylistic identity. [164 words]		

Table 3: Signal loss for prompts and corresponding generated images

References

[1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.