Controlled Image Editing Mechanisms Based on Diffusion Models

Master's Educational Program: Data Science

Student: Arina Chumachenko

Moscow 2024

Contents

Research Problem / Question(s)

The challenge of personalized image generation and editing is a cornerstone of advancements in controlled image editing mechanisms based on diffusion models. This task involves creating or modifying images such that they not only align with user-provided text prompts but also accurately represent a specific concept, such as an object, person, or style. Despite significant progress in generative modeling, several critical issues persist in achieving effective personalization.

First, modern text-to-image generative models struggle to balance fidelity to the user-provided concept with adherence to the textual prompt. Achieving this balance is crucial for generating high-quality images that faithfully combine the user's desired subject with contextual elements described in the text prompt. Models often fail in this regard, producing outputs that either misrepresent the concept or deviate from the textual instructions.

Second, these models are highly prone to overfitting on user-defined concepts during fine-tuning. Overfitting occurs because the model excessively adapts its attention mechanisms to the learnable concept, effectively ignoring broader contextual information in the text prompt. This phenomenon reduces the generalization capabilities of the model, limiting its ability to integrate new contexts or styles into the generated images.

Another significant issue is *semantic* (or *language*) *drift*, a phenomenon observed in the text space of models like CLIP after fine-tuning. Semantic drift alters the relationship between personalized concepts and their associated superclasses. For instance, as described in the study [статья ClassDiff], visualizing the CLIP text sample space before and after fine-tuning reveals a striking pattern: within a superclass of objects (e.g., "dogs"), the fine-tuned target concept (e.g., "a cute dog") diverges significantly from its superclass representation. This semantic shift indicates that the model's understanding of the target concept evolves during fine-tuning, potentially detaching it from the generalizable knowledge embedded in the superclass.

This divergence has practical implications. The loss of alignment with superclasses can degrade the compositional capabilities of generative models, as they struggle to integrate personalized concepts into new and complex scenarios. For example, a fine-tuned model may fail to accurately depict the personalized concept "a tall dog" if the semantic drift disrupts its understanding of the broader category "dogs."

Addressing these challenges requires innovative approaches to maintain semantic alignment, prevent overfitting, and balance prompt adherence with concept fidelity. Controlled image editing mechanisms based on diffusion models offer a promising direction to overcome these limitations. By incorporating techniques that preserve semantic relationships, enhance prompt compliance, and optimize fine-tuning strategies, the full potential of personalized image generation and editing can be unlocked.

Goals and objectives

The primary goal of this research is to develop methods for the correct integration of new concepts into the embedding space of text-to-image generative models to enhance the quality of output images. By achieving this, the research seeks to enable controlled modification of concepts, ensuring that the semantics of personalized phrases remain closely aligned with their corresponding category-centered words. This alignment is critical for preserving the compositional and semantic integrity of the generated images.

The research aims to learn effective text embeddings for new concepts that not only accurately represent the concepts but also seamlessly integrate with existing tokens in the model's language-vision dictionary. The enhanced dictionary would bind newly introduced words with specific user-defined subjects, enabling the model to synthesize novel, photorealistic images of these subjects in diverse contexts while maintaining their key identifying features. Ultimately, the research strives to expand the capabilities of generative models, allowing for precise personalization and editing without compromising image quality or semantic coherence.

The main objectives of this research include:

1. Conducting an in-depth literature review of current approaches to personalized image generation and editing, identifying their limitations, and exploring potential areas for improvement.
2. Building on baseline methods, the study seeks to propose an improved approach to personalized image generation, leveraging diffusion models for enhanced control and quality.
3. Introducing new regularization techniques to address challenges such as overfitting and semantic drift, ensuring that the model generates high-fidelity images that align with both the textual prompts and the personalized concepts.
4. Designing mechanisms to prevent semantic drift during fine-tuning by maintaining alignment between personalized concepts and their superclasses, thus preserving the generalization capabilities of the model.
5. Performing extensive experiments to validate the proposed methods, using quantitative and qualitative metrics to evaluate the quality, fidelity, and diversity of generated images.
6. Creating a comprehensive benchmark to compare the performance of existing and newly proposed methods, providing a standardized evaluation framework for personalized generative models.

By addressing these objectives, the research aims to push the boundaries of controlled image editing mechanisms, offering novel solutions to longstanding challenges in the field of generative modeling.

This work introduces a novel approach to addressing key challenges in personalized generative models, including semantic drift, overfitting, and the integration of new concepts into pre-trained text-to-image models. Unlike existing methods, this research focuses on combining advanced regularization techniques with a tailored fine-tuning strategy to ensure that new concepts maintain semantic coherence with their superclasses while integrating seamlessly into the model’s embedding space.

Additionally, the study proposes innovative solutions for enhancing model performance through controlled text embedding modifications, aiming to expand the model's language-vision dictionary without extensive retraining or compromising the fidelity of generated images. The introduction of a robust benchmark, along with new evaluation metrics tailored to personalized generation tasks, provides a significant contribution to the field, enabling the systematic assessment and comparison of different methods.

By offering these novel contributions, this research not only advances the state-of-the-art in controlled image editing but also paves the way for more effective and efficient personalization in generative models.

Literature Review

Generative models have transformed computer vision, making it possible to generate high-quality images from textual descriptions. Foundational methods have paved the way for advancements in this field, beginning with Denoising Diffusion Probabilistic Models (DDPMs) [1]. These models iteratively denoise Gaussian noise vectors to produce realistic images, guided by probabilistic models. Latent Diffusion Models (LDMs) [2] improve efficiency by operating in a compressed latent space, reducing computational overhead without sacrificing quality. Building on these, Denoising Diffusion Implicit Models (DDIMs) [3] introduce deterministic sampling for faster generation.

Stable Diffusion, a scalable variant of LDMs, incorporates pre-trained text encoders like CLIP [4] for robust text-to-image translation. CLIP, a dual encoder, aligns textual and visual representations, facilitating better multimodal understanding. These foundational models have set the stage for personalized image generation methods, addressing the demand for more specific and user-defined outputs.

The next part of the articles is devoted to personalized generation methods and two crucial methods in this area are Textual Inversion [5] and DreamBooth [6]. The main question that the authors of articles on personalized image generation ask is: how, having an image of a concept (for example, few images of your pet or favorite character) and a text prompt (for example, “in the jungle"), to get an image that alien well enough to this text prompt, and also conveys with high accuracy the concept.

The authors of the paper [5] set out to enable the generation of user-specified concepts guided by language. Their objective is to encode these concepts into an intermediate representation of a pre-trained text-to-image model. The study demonstrates that earlier approaches [7], which utilized embedding spaces shown to be expressive enough to capture basic image semantics, relied on contrastive or language completion objectives. However, these methods did not require a deep visual understanding of the image. This paper aims to discover pseudo-words capable of guiding the generative process. To achieve this, the authors propose finding such pseudo-words through a visual reconstruction objective, applying their method to LDMs.

In general, text encoder models start by processing input text. Each word or sub-word in the input string is converted into a token — an index from a pre-defined dictionary. Each token is then mapped to a unique embedding vector, which can be retrieved through an index-based lookup. These embedding vectors are typically learned as part of the text encoder.

In this study, the embedding space of the text encoder is chosen as the target for inversion. A placeholder string S^* is used to represent the new concept to be learned. The authors intervene in the embedding process by replacing the vector corresponding to the tokenized string with a new, learned embedding v^*, effectively "injecting" the concept into the model's vocabulary. Once integrated, the new embedding can be used within sentences, allowing the concept to be incorporated as if it were a standard word.

The core concept behind textual inversion involves discovering these new embeddings. To do so, the authors use 3–5 images representing the target concept in various settings, such as different backgrounds or poses. The new embedding v^* is obtained through direct optimization by minimizing the LDM loss function over the small set of sample images. To condition the generative process, the authors randomly generate neutral context prompts like “A photo of a small S^*” or “A rendition of S^*.”

However, this method suffers from a number of limitations: firstly, it has a weak accuracy of the reconstructed concepts, and secondly, this approach requires lengthy optimization times (for a single concept it takes roughly two hours).

Another main problem that personalization models face is *language drift*, a phenomenon that lies in the fact that the model forgets the broader class properties of the subject. The article [7] describes the DreamBooth method, which offers a solution to this problem.

This work introduces an innovative method for personalizing text-to-image diffusion models, enabling the generation of photorealistic images of a specific subject based on just 3-5 reference images, as it was in the previous method. The authors achieve this by fine-tuning a pre-trained model to associate a unique identifier with the subject, allowing it to generate novel images contextualized in diverse scenes while preserving the subject’s distinct features. This is achieved by fine-tuning all UNet model.

The core contribution lies in binding rare token identifiers to the subject and fine-tuning the model using a novel autogenous, class-specific prior preservation loss. This loss prevents overfitting and language drift, maintaining the model’s ability to generate diverse outputs for the subject’s class. By leveraging the model’s semantic prior, this approach facilitates generating variations of the subject in different poses, lighting, and environments, even those absent from the reference images.

The main problems that are being addressed in this work include language drift and reduced diversity. The proposed class-specific prior preservation loss mitigates the first problem by supervising the model with its own generated samples. The second challenge appears because of the fact that fine-tuning on limited data risks snapping to the reference images. The authors' technique encourages variability by leveraging class priors.

The authors provide an efficient prompt design process, labeling input images with a placeholder identifier and class noun (e.g., "a [V] dog"). They also use rare-token identifiers to minimize interference with existing model priors, ensuring effective personalization without semantic conflicts. Fine-tuning all layers of the model further enhances fidelity.

Despite its strengths, the limitations of DreamBooth include difficulty generating rare subject variations, entanglement between context and subject appearance and challenges with generating rare or complex subjects and contexts.

Another approach AttnDreamBooth [8] builds upon the foundational DreamBooth framework by incorporating enhanced attention mechanisms to achieve greater fidelity in personalized outputs. However, this method is limited by its extensive training time, making it less viable for time-sensitive applications. Furthermore, its consistent training steps across various concepts reduce its adaptability, making it less effective for a diverse range of personalization tasks. Despite these drawbacks, AttnDreamBooth sets a strong benchmark in improving output quality through refined attention mechanisms.

Custom Diffusion [9] offers a more efficient approach by fine-tuning specific components such as text embeddings and select parameters in the U-Net. This method achieves faster tuning for multi-concept customization, striking a balance between efficiency and output quality. Its lightweight design enhances flexibility across different concepts, though the reliance on parameter selection introduces potential trade-offs in fidelity when dealing with more complex or nuanced concepts.

SVDiff [10] introduces a highly compact and efficient approach to personalization by fine-tuning the singular values of weight matrices. This innovation reduces the risk of overfitting and mitigates issues like language drift. Additionally, SVDiff employs a Cut-Mix-Unmix data augmentation technique to improve the generation of multi-subject images, complemented by a straightforward text-based image editing framework. With a model size of just 1.7MB, SVDiff stands out as a practical solution compared to much larger alternatives like DreamBooth (3.66GB) and Custom Diffusion (73MB). However, the method's reliance on singular value manipulation may limit its adaptability in handling highly diverse or intricate concepts.

XTI [11] expands the expressiveness of text embedding spaces by utilizing multiple tokens, with each token assigned to a specific attention layer. This approach provides granular control over concept representation, improving the model's ability to customize outputs in detail. However, the increased complexity associated with managing multiple tokens introduces additional computational overhead, potentially reducing its applicability in scenarios where lightweight processing is essential.

NeTI [12] advances personalization by incorporating dependencies on denoising timesteps and U-Net layers, thereby expanding the text embedding space. This enhancement enables finer temporal and spatial control, improving both coherence and detail in generated images. While NeTI excels in achieving high fidelity for complex tasks, its increased training complexity poses scalability challenges, particularly in real-world applications requiring efficiency and adaptability.

The authors of the article [13] propose Prompt-Aligned Personalization (PALP) to address the challenge to balance adherence to user-defined prompts with the accurate representation of personalized subjects. PALP focuses on single-prompt alignment and emphasizes optimizing for a specific textual prompt. For instance, generating "a sketch of [my cat] in Paris" not only involves representing the subject faithfully but also maintaining the stylistic and contextual elements.

This method combines personalization, where the model learns a subject from a few or even a single image, and prompt alignment, which ensures the generated image adheres to the target text prompt. This dual optimization is achieved by fine-tuning the model while incorporating score distillation sampling (SDS) [2, 14, 15, 16] to guide predictions.

Overfitting to a small subject dataset often results in outputs dominated by training images, ignoring the prompt. PALP addresses this by steering model predictions towards prompt-aligned outputs during training. It ensures the model captures the defining features of the subject without reconstructing irrelevant details like backgrounds. Fine-tuning with Low-Rank Adaptation (LoRA) [17] updates only a subset of network weights, improving efficiency. This method leverages pre-trained model knowledge to maintain alignment with the target prompt, sidestepping the need for additional large-scale datasets. PALP also supports multi-subject scenarios and inspiration from artistic references, broadening its applicability. The paper demonstrates PALP’s advantage over baselines in both single- and multi-shot settings. It generates highly prompt-compliant images without requiring extensive pre-training.

The authors introduce prompt-aligned score sampling, which adjusts the model’s noise predictions during the backward diffusion process to align outputs with the target prompt. This prevents overfitting while ensuring text-alignment. Experiments reveal that balancing guidance scales between personalization and text-alignment branches significantly enhances results.

PALP extends the capabilities of text-to-image models, enabling: personalization with a single reference image and complex scene generation incorporating multiple subjects. However, the method requires re-personalization for new prompts, limiting real-time applications.

PALP emphasizes balancing adherence to textual prompts with accurate subject representation. It employs dual optimization, combining personalization with prompt alignment through SDS. Fine-tuning is achieved using LoRA for efficient adjustments. PALP is effective for both single- and multi-subject scenarios, supporting complex scene generation. However, its reliance on re-personalization for new prompts limits its applicability in real-time scenarios.

However, it is worth noting that all previous methods mainly solved the problem of language drift and, in general, the task of personalized generation from the UNet side, but what if we approach the personalization task on the text encoder side, and in particular its fine-tuning. This task was set by the authors of the article [18], who presented the TestBoost method.

Recent advancements in text-to-image models have enabled personalized image generation from natural language prompts. However, these methods often falter with a single reference image, overfitting the input and producing repetitive outputs. The paper [18] introduces TextBoost, a novel approach that exclusively fine-tunes the text encoder to mitigate overfitting, achieving high-quality one-shot personalization.

TextBoost focuses solely on the text encoder, unlike prior methods that modify the image generation module, so it provides selective fine-tuning. In addition, this method solves the problem that arose in Text Inversion, namely, it is economical in terms of memory and data storage. Both of these advantages are achieved through the introduction of three key techniques to enhance personalization performance.

Firstly, the authors notice that the models suffer from augmentation leaking [19], due to the fact that after fine-tuning of the text encoder, it is difficult for models to separate the original subject from the background, which requires an increase in the amount of data for training. To mitigate this leaking and manually adjust the model, the authors suggest additionally using *augmentation tokens*, which help to enhance feature disentanglement, separating subject-relevant features from irrelevant details, and to reduce overfitting of the model by automatically mapping transformations (e.g. vertical flip, rotation by 90 degrees, etc.) to the augmentation token A^*. For instance, a sample image of a dog rotated by 90 degrees would be accompanied by the prompt ' A^* photo of V^*'.

Secondly, on a par with previous methods the authors use specific loss to prevent language drift [20], maintaining generalization across diverse prompts. They use knowledge-preservation loss, which is essentially a cosine similarity between text embeddings from the pre-trained text encoder and the online text encoder. Thus, this loss ensures that the online text encoder retains its prior knowledge.

Lastly, the authors notice that the impact of text prompts on the neural network output is proportional to the input's noise level. So they propose a timestep sampling method based on the signal-to-noise ratio (SNR) [21] of noisy input. It helps to improve training efficiency by prioritizing noise levels.

TextBoost introduces a practical solution for one-shot text-to-image personalization, balancing quality, efficiency, and creative control. But this method also still suffers from the quality of the resulting images.

Like the authors of previous articles, the authors of the article [22] note that fine-tuning often leads to overfitting, causing a loss of compositional ability, and they also identify the root cause as semantic drift, where the fine-tuned concept diverges from its superclass. To address this, they propose ClassDiffusion, a technique incorporating semantic preservation loss (SPL) to maintain the compositional capabilities of fine-tuned models.

The paper highlights how personalization tuning shifts the target concept away from its superclass in the CLIP text space, leading to reduced cross-attention strength and weaker compositional capabilities. Their specific SPL minimizes the cosine distance between text embeddings of personalized and superclass phrases, preserving the semantic alignment and restoring compositional ability. Moreover, recognizing the limitations of existing CLIP-T metrics, the authors introduce BLIP2-T, a more effective and fair evaluation metric for text-image alignment.

Empirical analysis of ClassDiffusion showed that visualizing CLIP text space and cross-attention layers revealed that semantic drift reduces entropy in compositional probability, impairing the model's ability to combine target concepts with other elements. Nevertheless this approach has some limitations, such as: an application to fine-grained human face generation remains unexplored and selecting appropriate center words for objects with mixed categories requires manual experimentation.

The proposed in the article [23] Context Regularization (CoRe) method addresses the issue of balancing identity preservation with prompt-aligned text-image generation by improving how new concepts integrate into the embedding space of text encoders, particularly in CLIP-based models.

Generating accurate, prompt-aligned images depends on a robust semantic understanding of both the new concept and its interactions with surrounding context tokens. CoRe focuses on regularizing the embeddings and attention maps of context tokens to ensure consistent behavior when integrating a new concept. This approach involves two main innovations: embedding regularization and attention regularization.

The embedding regularization is essentially a cosine similarity between text embeddings of received after text encoder prompts: the first one contains a superclass token and received from prompt “A photo of a dog”, and the second one contains a placeholder token and received from prompt “A photo of a V^*”. Therefore, by comparing context tokens in prompts containing the new concept with a reference prompt using a superclass token, CoRe minimizes semantic drift.

In the attention regularization CoRe enforces consistency in cross-attention maps, ensuring context tokens behave predictably even as new concepts are introduced. Thus, in this case, the embeddings related to the superclass and the embeddings related to the placeholder, received after the text encoder, are sent to UNet and, after passing through 16 of UNet layers, 16 pairs of attention maps are obtained at the output, after that they are averaged and their mean squared error is calculated.

CoRe employs a similarity constraint between output embeddings and attention maps of context tokens for prompts with new and reference concepts. The regularization is performed without requiring image generation, enhancing the method's generalization. Additionally, embedding rescaling mitigates optimization issues that could harm text alignment.

A two-stage training strategy further refines CoRe: at the first stage the model learns a compatible text embedding using CoRe, at the second stage the fine-tuning of U-Net layers takes place to precisely identify the concept. This method also functions as a test-time optimization tool for enhancing generation quality. However, challenges persist in handling complex compositions, particularly with pre-trained model limitations.

Evaluating personalized image generation methods relies on metrics like CLIP text and image similarity, which assess semantic alignment between text and generated images, and DINO image similarity, which measures visual fidelity to reference images.

Personalized image generation has evolved significantly, with methods ranging from foundational approaches like Textual Inversion to advanced techniques such as CoRe. While these methods address various challenges, common limitations persist, including semantic drift, overfitting, and efficiency constraints. Future research should prioritize real-time personalization, complex compositional generation, and the refinement of evaluation metrics to further advance the field.

Methods

To enable high-quality, text-aligned image generation, our method based on the article [23] builds on Context Regularization to learn precise and semantically consistent text embeddings for new concepts. The goal is to ensure that these embeddings integrate seamlessly with the pre-trained token space, preserving the compositional and semantic interactions required for accurate text-to-image synthesis.

Proper output embeddings of context tokens depend on the accurate learning of the new concept's input embedding. Misalignment in the new embedding can adversely affect the semantics of the entire prompt.

When substituting a new concept token with another, the embeddings and attention maps of the surrounding context tokens should remain consistent. Misalignment, often observed in overfitted embeddings, disrupts this consistency.

To address these challenges, the method employs regularization strategies to align the new concept embedding with its super-category, thereby preserving the integrity of the surrounding context.

The first regularisation is embedding regularization. We regularize the embeddings of context tokens by creating two prompts: one with the new concept token and another with its super-category token. The similarity constraint ensures alignment between the two sets of output embeddings:

$$L_{emb} = \frac{1}{n-1} \sum_{i=1,\, i \neq k}^{n} \left[1 - \cos\left(E(v_i),\ E(v_i')\right)\right],$$

where $E(v_i)$ and $E(v_i')$ are the output embeddings of the context tokens for the prompts with the new concept and super-category, respectively.

The second regularisation is attention regularization. Attention maps generated for context tokens are regularized to ensure that the introduction of a new concept token does not distort their distribution. We minimize the squared difference between the mean values of attention maps for the two prompts:

$$L_{attn} = \frac{1}{n-1} \sum_{i=1,\, i \neq k}^{n} \left[\mu\left(M_i^{1:16}\right) - \mu\left(M_i^{'1:16}\right)\right]^2,$$

where $M_i^{1:16}$ denotes the mean across attention maps for the context tokens.

Also we can add pairwise loss regularization. During the training phase, after passing the text encoder, we compute the encoder's hidden states for prompts containing the new concept token and the placeholder token. For each, we calculate pairwise cosine similarities between tokens, yielding two matrices. The absolute difference of these matrices forms the pairwise loss, which is added to the original diffusion loss to maintain consistency between placeholder and subclass interactions.

The next type of regularization is Hinge regularization. To address semantic drift, we compute cosine similarities between the embeddings of the placeholder tokens and those of their super-category. A hinge loss with a predefined margin enforces alignment, ensuring that the placeholder tokens retain their semantic relation to the superclass throughout training.

To counteract the excessively large scales of new concept embeddings during optimization, we rescale their norms to match the scale of the previous optimization step:

$$v_s^* = \frac{v_s^*}{\left\|v_s^*\right\|}\left\|v_{s-1}^*\right\|$$

where s denotes the optimization step. This technique preserves semantic alignment without degrading concept identity.

All of the above applies to the first stage of model training, which is editable embedding learning. The method is applied to learn a text embedding compatible with the model’s token space, resulting in an editable representation of the new concept.

The second stage is fine-tuning: the text embedding is frozen, and the U-Net is fine-tuned to capture the unique identity of the new concept while maintaining compatibility with the surrounding tokens.

Despite its advantages, this approach has limitations:

- The alignment of new concepts with their super-categories may still result in subtle drift, particularly for abstract or highly diverse categories.
- The multi-stage training and regularization mechanisms significantly increase the training time and resource requirements.

- The two-stage strategy may occasionally lead to over-optimization in the second stage, where excessive fine-tuning compromises the editability of the learned embeddings.

These limitations highlight areas for further refinement, including the development of more efficient regularization methods and strategies to balance identity preservation with computational efficiency.

LIST OF REFERENCES

[1] Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models //Advances in neural information processing systems. – 2020. – T. 33. – C. 6840-6851.

[2] Rombach R. et al. High-resolution image synthesis with latent diffusion models //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2022. – C. 10684-10695.

[3] Song J., Meng C., Ermon S. Denoising diffusion implicit models //arXiv preprint arXiv:2010.02502. – 2020.

[4] Radford A. et al. Learning transferable visual models from natural language supervision //International conference on machine learning. – PMLR, 2021. – C. 8748-8763.

[5] Gal R. et al. An image is worth one word: Personalizing text-to-image generation using textual inversion //arXiv preprint arXiv:2208.01618. – 2022.

[6] Ruiz N. et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2023. – C. 22500-22510.

[7] Cohen N. et al. "This is my unicorn, Fluffy": Personalizing frozen vision-language representations //European conference on computer vision. – Cham : Springer Nature Switzerland, 2022. – C. 558-577.

[8] Pang L. et al. AttnDreamBooth: Towards Text-Aligned Personalized Text-to-Image Generation //arXiv preprint arXiv:2406.05000. – 2024.

[9] Kumari N. et al. Multi-concept customization of text-to-image diffusion //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2023. – C. 1931-1941.

[10] Han L. et al. Svdiff: Compact parameter space for diffusion fine-tuning //Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2023. – C. 7323-7334.

[11] Voynov A. et al. p+: Extended textual conditioning in text-to-image generation //arXiv preprint arXiv:2303.09522. – 2023.

[12] Alaluf Y. et al. A neural space-time representation for text-to-image personalization //ACM Transactions on Graphics (TOG). – 2023. – T. 42. – №. 6. – C. 1-10.

[13] Arar M. et al. PALP: prompt aligned personalization of text-to-image models //SIGGRAPH Asia 2024 Conference Papers. – 2024. – C. 1-11.

[14] Saharia C. et al. Photorealistic text-to-image diffusion models with deep language understanding //Advances in neural information processing systems. – 2022. – T. 35. – C. 36479-36494.

[15] Poole B. et al. Dreamfusion: Text-to-3d using 2d diffusion //arXiv preprint arXiv:2209.14988. – 2022.

[16] Katzir O. et al. Noise-free score distillation //arXiv preprint arXiv:2310.17590. – 2023.

[17] Hu E. J. et al. Lora: Low-rank adaptation of large language models //arXiv preprint arXiv:2106.09685. – 2021.

[18] Park N. H., Kim K., Shim H. TextBoost: Towards One-Shot Personalization of Text-to-Image Models via Fine-tuning Text Encoder //arXiv preprint arXiv:2409.08248. – 2024.

[19] Karras T. et al. Training generative adversarial networks with limited data //Advances in neural information processing systems. – 2020. – T. 33. – C. 12104-12114.

[20] Lee J., Cho K., Kiela D. Countering language drift via visual grounding //arXiv preprint arXiv:1909.04499. – 2019.

[21] Johnson D. H. Signal-to-noise ratio //Scholarpedia. – 2006. – T. 1. – №. 12. – C. 2088.

[22] Huang J. et al. ClassDiffusion: More Aligned Personalization Tuning with Explicit Class Guidance //arXiv preprint arXiv:2405.17532. – 2024.

[23] Wu F. et al. CoRe: Context-Regularized Text Embedding Learning for Text-to-Image Personalization //arXiv preprint arXiv:2408.15914. – 2024.