Mitigating Semantic Collapse in Generative Personalization with a Surprisingly Simple Test-Time Embedding Adjustment

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we investigate the semantic collapsing problem in generative personalization, an under-explored topic where the learned visual concept (V^*) gradually shifts from its original textual meaning and comes to dominate other concepts in multi-concept input prompts. This issue not only reduces the semantic richness of complex input prompts like "a photo of V^* wearing glasses and playing guitar" into simpler, less contextually rich forms such as "a photo of V^*" but also leads to simplified output images that fail to capture the intended concept. We identify the root cause as unconstrained optimisation, which allows the learned embedding V^* to drift arbitrarily in the embedding space, both in direction and magnitude. To address this, we propose a simple yet effective training-free method that adjusts the magnitude and direction of pre-trained embedding at inference time, effectively mitigating the semantic collapsing problem. Our method is broadly applicable across different personalization methods and demonstrates significant improvements in text-image alignment in diverse use cases. Our code is anonymously published at https://anonymous.4open.science/r/Embedding-Adjustment.

1 Introduction

Text-to-image (T2I) diffusion models have achieved unprecedented fidelity and flexibility in image generation and sparked growing interest in generative personalization [9, 25]. This emerging problem aims to generate images of a specific user-defined visual concept (e.g. a particular person, pet, or object) in different contexts (e.g. on the beach) using a small set of user-provided reference images paired with text prompts describing the desired context. The core objective is to generate visually compelling images that faithfully preserve the unique characteristics of the personal concept while remaining semantically aligned with the textual prompt. Despite recent progress, misalignment between the generated image and the textual prompt is still a major concern.

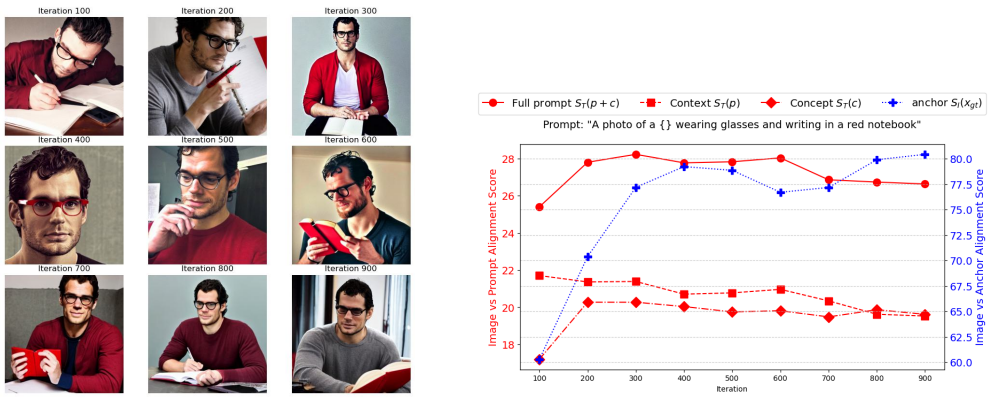
A robust generative personalization method should allow the user-defined visual concept to be composed with arbitrary contexts in the text prompt without losing fidelity or expressiveness. However, existing approaches often struggle to maintain prompt and generated image alignment, particularly with complex or multi-concept prompts [15, 37]. In these scenarios, the introduced concept can overpower or distort other elements in the prompt, leading to unsatisfactory generations. This issue has commonly been attributed to *language drift* or overfitting [25], arising from optimising on a limited number of reference images. Beyond overfitting, other factors have also been attributed, such as the limited expressiveness of textual embeddings which compress complex visual concepts into single tokens [36, 20] and the entangled nature of reference sets, where samples may contain co-occurring objects or irrelevant contextual features [1, 14].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In this work, we investigate an under-explored failure mode of generative personalization, which we term the *semantic collapsing problem* (SCP). SCP refers to the phenomenon where the learned personalized token for a new concept loses its original textual semantic meaning while acquiring increased visual information from the reference concept during the personalization process. As a result, when a user combines the personalized token with other descriptive context in a prompt, the model's output is overwhelmingly biased towards the personalized concept, often neglecting the other intended elements in the prompt. For example, if one learns a token V^* to represent a particular dog, a prompt like "V^* playing with a ball in a park" may yield an image that vividly depicts the dog but fails to properly render the ball or the park background. In essence, the prompt's semantic richness collapses to a simpler form centred on V^*. While SCP does not severely affect trivial prompts (e.g., "a photo of V^*", where the entire prompt is just the concept itself), it undermines the compositionality of the T2I diffusion model on complex prompts. Notably, SCP is distinct from the language drift problem, which describes a different failure where the personalized model overfits to a learned concept and loses generalisation. In SCP, the issue arises when the personalized embedding no longer retains any meaningful textual semantics.

We identify the root cause as *unconstrained optimisation*, which allows the learned embedding to drift arbitrarily in the embedding space, both in direction and magnitude. We propose a simple yet effective remedy: a training-free, test-time embedding adjustment strategy that realigns the learned concept embedding with its original semantic meaning at inference time. The key idea is to calibrate the embedding's magnitude and direction to be closer to that of its reference concept. This adjustment is done without altering the model weights or requiring any additional training. The embedding is modified on-the-fly before image generation. By enforcing a small rotation and rescaling in the text encoder's latent space, we constrain the personalized token to behave more like a regular word, ensuring that it contributes to the image generation in balance with other tokens in the prompt. Our approach is lightweight and broadly compatible with different personalization methods and demonstrates significant improvements in text-image alignment in diverse use cases.

In summary, our contributions are as follows: ❶ We define the semantic collapsing problem (SCP) in generative personalization problem, and provide an empirical analysis of its existence in both textual and image spaces. Our analysis reveals its root cause, which is the unconstrained optimisation. ❷ We propose the test-time embedding adjustment, a novel solution that requires no additional training, to mitigate SCP by aligning the learned embedding's direction and norm with its original semantic concept. ❸ We demonstrate that our proposed approach can be applied into different personalization methods such as Textual Inversion [9], DreamBooth [25], and Custom Diffusion [16] to significantly improve image generation in complex prompts across a wide range of scenarios.



(a) Intermediate generated images

(b) Alignment scores

Figure 1: Illustration of the SCP on Textual Inversion. Left: The intermediate generated images $\hat{x}$ of a prompt 'a photo of a V^* wearing glasses and writing in a red notebook'. The generated image is gradually biased towards the personalized concept V^* (i.e., easier to recognize as 'Henry Cavill') and loses the context p (i.e., harder to recognize as 'writing in a red notebook') through out the personalization process. Right: The alignment scores (average over 100 random seeds) which empirically validate the SCP. The alignment $S(\hat{x}, p)$ ($\square$) with the context p drops over time, while the alignment $S(\hat{x}, x_{gt})$ ($+$) with the ground truth x_{gt} increases.

2 Personalizing T2I Generative Models

Given a text-to-image diffusion model ϵ_θ, where $\epsilon_\theta(x_t, t, p)$ represents the predicted noise at time step t given the textual embedding $\tau(p)$ of a prompt p and the noisy intermediate vector x_t [11, 29, 24]. And given a set of personal images $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ and a pre-trained T2I model ϵ_θ, the goal of generative personalization is to identify a textual embedding v^* associated with a specific verbalizable keyword V^* (e.g., 'sks', '<new>', etc.). This keyword represents the implicit visual concept shared in the reference set $\mathcal{X}$, enabling the model to generate images with the personal concept using any textual prompt p containing the keyword V^*, e.g., $\lfloor p, V^* \rfloor$ = 'A photo of V^* playing on a beach', where $\lfloor ., . \rfloor$ is the sentence construction operator.

This keyword represents the implicit visual concept shared in the reference set $\mathcal{X}$, enabling the model to generate images with the personal concept using any textual prompt p containing the keyword V^*, e.g., $\lfloor p, V^* \rfloor$ = 'A photo of V^* playing on a beach', where $\lfloor ., . \rfloor$ is the sentence construction operator. We denote M is the embedding matrix of the entire vocabulary of the text encoder τ, and M_k is a specific row of the matrix corresponding to the token $k \in \text{vocab}_\tau$, where k can be V^* (i.e., $v^* = M_{V^*}$) or any arbitrary token like 'dog', 'cat', etc. We denote $\hat{x} = G(p)$ as the image generation model that takes a prompt p as input and outputs an image $\hat{x}$. We discuss two representative methods for personalizing T2I generative models: Textual Inversion (TI) [9] and DreamBooth [25] and other related works in Appendix A.

3 Semantic Collapsing Problem in Generative Personalization

3.1 Semantic Collapsing Problem

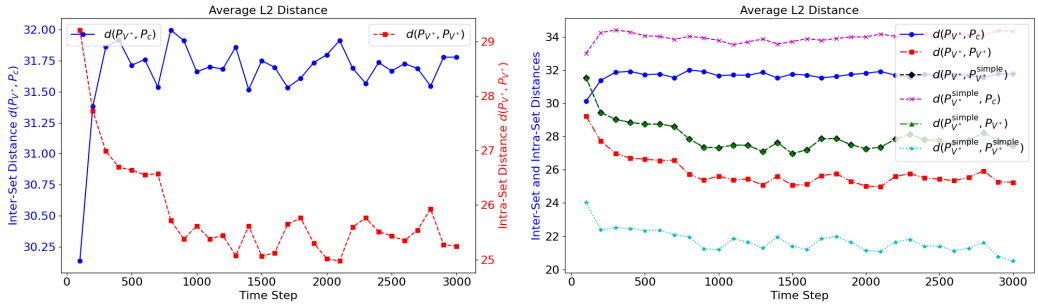
Problem Statement. The **semantic collapsing problem** (SCP) refers to the phenomenon where the keyword V^* loses its original *textual semantic meaning* while acquiring increased *visual information* from the reference concept during the personalization process. As a result, a prompt $\lfloor p, V^* \rfloor$, consisting of a context p and the concept V^*, becomes dominated by the learned concept V^*, eventually collapsing to a simplified form.

Why SCP Matters. We argue that SCP may not pose a serious issue for simple prompts, e.g., $\lfloor p, V^* \rfloor$ = 'a photo of V^*', where the primary information conveyed is still the visual concept V^*. However, for more complex prompts where the context p contributes meaningfully to the overall semantics, such as 'a photo of V^* wearing glasses and writing in a red notebook', SCP becomes more problematic as illustrated in Figure 1. In such cases, the generated image is more likely to be dominated by the learned concept V^* and less likely to reflect the intended context p.

Comparison with Other Challenges in Generative personalization. First, we emphasise that SCP is not specific to the two representative methods we study (TI and DB), but is a general issue in personalization. Second, SCP is distinct from other recognised challenges (ref. Section A) such as the *language drift* problem [25], which describes how the personalized model ϵ_θ overfits to a learned concept and loses generalisation, for instance, generating a personalized dog image from a prompt like 'a photo of a dog', even when no personalized keyword is present. In contrast, SCP refers to the case where the keyword V^* no longer retains any meaningful textual semantics, instead encoding only visual information. Third, while SCP contributes to the broader challenge of misalignment between generated images and prompts, a major concern in generative personalization, it stems from a specific cause: the unconstrained optimisation of the embedding during personalization, which has not been thoroughly studied in prior work.

3.2 Empirical Hunting for SCP

In this section, we present empirical evidence supporting the existence of the semantic collapsing problem and its impact on generation quality. Our key findings are as follows: ❶ **Existence of SCP.** SCP exists in the textual domain, where the prompt $\lfloor p, V^* \rfloor$ is dominated by the learned embedding V^* and the semantic meaning of the entire prompt gradually collapses to the learned embedding V^*, i.e., $\tau(\lfloor p, V^* \rfloor) \rightarrow \tau(V^*)$. ❷ **Negative Impact on Generation Quality.** SCP leads to the degradation/misalignment in generation quality in the image space, i.e., $G(\lfloor p, V^* \rfloor) \rightarrow G(V^*)$, particularly for prompts with complex semantic structures. ❸ **Surprisingly Positive Impact.** SCP can also lead to the positive impact on generation quality, particularly for prompts where the concept



(a) (b)

Figure 2: (a) The inter-set distance $d(P_{V^*}, P_c)$ and intra-set distance $d(P_{V^*}, P_{V^*})$ over the personalization process, and (b) The distance between all possible pairs of sets, notably $d(P_{V^*}, P^{\text{simple}}_{V^*})$.

c requires a strong visual presence to be recognisable. ❹ **Root Cause of SCP.** SCP arises from unconstrained optimisation during personalization, which leads to arbitrary shifts (both in magnitude and direction) in the embedding of V^* away from its original semantic concept c.

3.2.1 Empirical Evidence for SCP in Textual Space

Recall our hypothesis: a keyword V^* initialised from a concept c to capture a visual target v_{gt} will lose its semantic meaning and dominate any arbitrary context p when combined into a prompt. A straightforward way to evaluate this *semantic shift* is to compare the embedding vectors M_{V^*} and M_c. However, modern LLMs and Diffusion Models use contextualised text embeddings, in which the surrounding context significantly influences the final representation of a prompt.

To address this, we propose measuring the difference between V^* and c in the presence of a diverse set of contextual prompts $A = \{a_1, a_2, \cdots, a_n\}$. These are generated by querying an LLM with the instruction: "Write 200 sentences with diverse topics and contents. Each sentence should be 10–30 words long and must include the keyword c." Sample sentences are provided in Table 3, and the full dataset and prompts can be found in the anonymous repository.

We then construct two sets of prompts: $P_{V^*} = \{\lfloor a_i, V^* \rfloor\}_{a_i \in A}$ and $P_c = \{\lfloor a_i, c \rfloor\}_{a_i \in A}$, which are used to assess the contextualised difference between V^* and c.

To quantify this difference, we compute distances between the sets P_{V^*} and P_c using four metrics: Euclidean, Hausdorff, Mahalanobis, and KL divergence. We also evaluate intra-set variability $d(P_{V^*}, P_{V^*})$ measuring the separation among items within P_{V^*}. The distance metrics are summarised in Table 5. We use Textual Inversion (TI) and DreamBooth (DB) to learn a personalized human face concept. The training data comprises 16 images from the CelebA dataset [19] (subject ID: 342, 'Henry Cavill'). The embedding v^* is initialised using the concept $c =$ 'man' for TI, and $c =$ 'a photo of a sks man' for DB.

Results. Figure 2a shows the average inter-set and intra-set distances, $d(P_{V^*}, P_c)$ and $d(P_{V^*}, P_{V^*})$, measured over training iterations. It can be seen that the inter-set distance $d(P_{V^*}, P_c)$ increases steadily over time, indicating that the learned embedding v^* progressively diverges from its initial textual semantic meaning c.

Interestingly, the intra-set distance $d(P_{V^*}, P_{V^*})$ decreases over time, suggesting that the embeddings of prompts $\lfloor a_i, V^* \rfloor$ within P_{V^*} become less diverse and more similar to one another. This reflects a growing dominance of the learned embedding v^* across prompts, effectively overriding the contextual variations in a_i and becoming the principal semantic component of each prompt.

To further verify this dominance effect, we introduce an additional set of simple prompts, denoted $P^{\text{simple}}_{V^*}$, which consists of 200 concise sentences such as 'a photo of a V^*', 'a portrait of a V^*', etc., where V^* is clearly the central concept. As shown in Figure 2b, the distance $d(P_{V^*}, P^{\text{simple}}_{V^*})$, which captures the difference between complex prompts $\lfloor p, V^* \rfloor$ and simple prompts $\lfloor V^* \rfloor$, decreases over time. This trend indicates that the representations of complex prompts become increasingly similar to those of simple prompts in P^{simple}, further supporting the hypothesis that v^* gradually dominates and collapses the semantic contribution of contextual components.

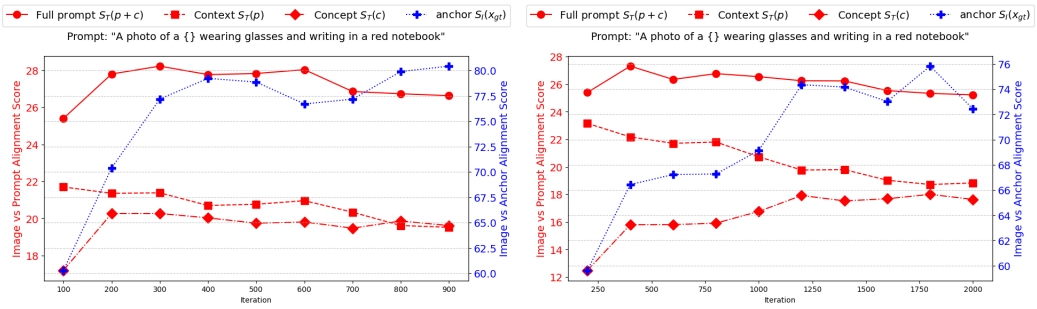


Figure 3: Analysis of the SCP on TI (left) and DB (right). Alignment with the ground-truth image ($S(\hat{x}, x_{gt}) - \Diamond$) increases over time, while alignment with the contextual part ($S(\hat{x}, p) - \square$) decreases.

3.2.2 The Two-Way Impacts on Personalization

In the previous section, we demonstrated the semantic collapsing problem in the textual space, where the prompt $\lfloor p, V^* \rfloor$ becomes increasingly dominated by the learned embedding V^*, causing the semantic meaning of the entire prompt to collapse toward V^*. In this section, we extend our analysis to investigate how this phenomenon impacts image generation quality.

Specifically, we generate personalized images $\hat{x} = G(\lfloor p, V^* \rfloor)$ using a list of prompts P (e.g., 'a photo of a V^* man holding a cat'), with 100 images generated per prompt. We evaluate the generation quality using the CLIP-Image-Image alignment score $S_I = S(\hat{x}, x_{gt})$, which measures the similarity between the generated image $\hat{x}$ and the ground-truth image x_{gt} of the reference concept. In addition, we compute three CLIP-Text-Image alignment scores: **(i)** CLIP$_T^p$ or $S_T^p = S(\hat{x}, p)$ alignment with the contextual part p (e.g., 'holding a cat'), and **(ii)** CLIP$_T^c$ or $S_T^c = S(\hat{x}, c)$ alignment with the original concept c (e.g., 'a man'). **(iii)** CLIP$_T^f$ or $S_T^f = S(\hat{x}, \lfloor p, c \rfloor)$ alignment with the full prompt.

Interestingly, we observe that the SCP has both negative and positive effects on generation quality, depending on the nature of the prompt. Our key findings (illustrated in Figure 3) are as follows:

(Unsurprising) The image-to-image alignment score S_I increases over time ($+$ line), indicating that V^* effectively captures the visual appearance of the target concept. This confirms that the learned embedding V^* successfully personalizes the visual identity from the reference set.

(Surprising Negative Impact) The context-to-image alignment score S_T^p decreases over time ($\square$ line), showing that generated images increasingly lose alignment with the contextual component p. This highlights the negative impact of semantic collapsing: as V^* dominates the prompt, the image generator pays less attention to the surrounding context.

(Even More Surprising Positive Impact) Surprisingly, for some prompts, the concept-to-image alignment score S_T^c increases over time ($\Diamond$ line). This suggests that at the early stages of training, when using the keyword c (e.g., 'man') in the prompt, the generated images may focus more on the context p rather than clearly representing c. As training progresses, however, the learned embedding V^* begins to dominate the prompt, causing the generator to produce images that better represent the core concept c and less of the context p. In such cases, semantic dominance enhances the generation of the intended subject. We find this positive effect tends to occur with prompts where the concept c requires a strong visual presence to be recognisable, such as $c =$ 'man' and $p =$ 'writing in a red notebook'. In these cases, generating a clear representation of the personalized subject (e.g., 'Henry Cavill') often results in close-up facial imagery, which inherently reduces the visibility of the context (e.g., the act of 'cycling'). Examples of this effect are shown in Figure 1a.

3.2.3 The Root Cause of SCP

In the previous sections, we demonstrated the semantic collapsing problem in both textual and image generation spaces. In this section, we investigate the underlying cause of this phenomenon. Our key findings are summarised below:

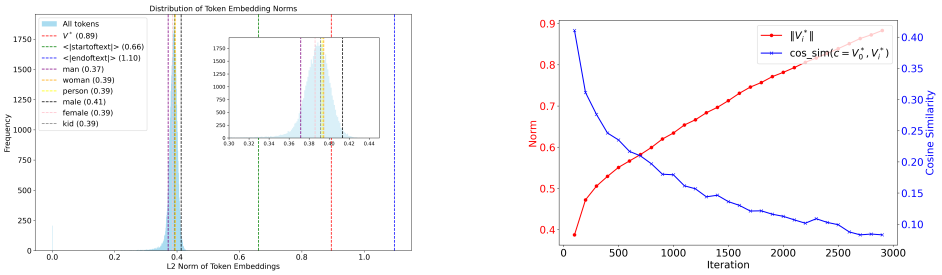


Figure 4: Left: The distribution of the norm of the token embedding M including special token V^*, Right: The semantic drift of V^* in term of magnitude and direction over time. The same phenomenon is observed in DreamBooth as shown in Figure 14.

① The semantic collapsing problem arises from the unconstrained optimisation process used in Equation 4 and Equation 5. Without any regularisation, the learned embedding V^* can deviate significantly from the original semantic meaning c in both magnitude and direction. Specifically, the embedding norm becomes much larger ($|M_{V^*}| \gg |M_c|$), and the cosine similarity between the two drops sharply ($\cos(M_{V^*}, M_c) \ll 1$), leading to a semantic drift.

② As a result of this semantic shift in V^*, the embedding of the entire prompt $\lfloor p, V^* \rfloor$ is also affected. That is, the prompt embedding becomes nearly identical to that of V^*, i.e., $\tau(\lfloor p, V^* \rfloor) \approx \tau(V^*) \neq \tau(\lfloor p, c \rfloor)$, which directly manifests the semantic collapsing problem discussed earlier.

To support this analysis, Figure 4 presents a histogram of embedding norms for all vocabulary tokens, along with the norm of V^* tracked over the course of optimisation. It is evident that the norm of V^* grows significantly, placing it in the long tail of the distribution—substantially larger than standard tokens such as 'man', 'woman', and 'person', and approaching that of special tokens like '<|startoftext|>' or '<|endoftext|>'.

It is worth noting that in some DreamBooth-based implementations (e.g., DreamBooth with LoRA in Diffusers [12]), the full text encoder is fine-tuned to incorporate the personalized concept, instead of updating a dedicated token embedding V^* as done in Textual Inversion. In such cases, while the individual embedding vector M is not explicitly altered, the semantic shift still occurs at the prompt level, i.e., $\tau(\lfloor p, V^* \rfloor) \approx \tau(V^*) \neq \tau(\lfloor p, c \rfloor)$.

Another worth mentioning is that these DreamBooth implementations already include a gradient clipping mechanism [12] to constrain parameter updates. However, this method was not designed with semantic stability in mind, and in practice, it does not prevent cumulative semantic drift. Even when the gradient norm is bounded, the embedding can still gradually shift over successive iterations.

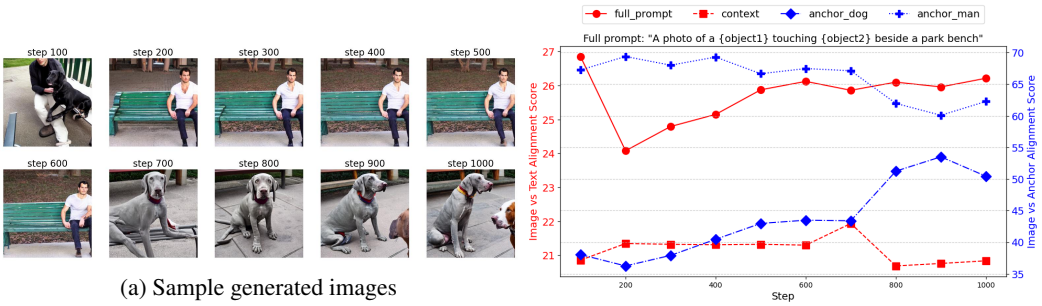
To the best of our knowledge, our work is the first to identify and explain the root cause of the semantic collapsing problem as a consequence of unregularised embedding dynamics.

3.3 SCP on Multi-Concepts Personalization

In this section, we investigate the question: *What is the SCP in the context of multi-concept personalization?* For instance, consider a prompt like 'A photo of a V^*_{man} touching a V^*_{dog} beside a park bench', where V^*_{man} and V^*_{dog} represent two independently personalized concepts. Will one concept dominate the other, as observed in single-concept personalization in the previous section?

To explore this, we conduct an experiment where the two concepts, V^*_{dog} and V^*_{man} (subject 342), are learned independently using Textual Inversion. We then construct a list of prompts that combine these two personalized concepts with an additional complex context, such as 'A photo of a V^*_{man} touching a V^*_{dog} beside a park bench' (refer to Table 4 for more examples).

Figure 5 illustrates the generated images where the embedding of V^*_{man} is held fixed, while the embedding of $V^*_{dog,step}$ is varied across different training steps. At early steps, $V^*_{dog,step}$ remains close to the original, generic 'dog' concept, while at later steps, it progressively captures more personalized visual information of the specific dog. This design allows us to observe how changes in the V^*_{dog} embedding influence the generation of V^*_{man} within the same prompt.



(a) Sample generated images

(b) Alignment scores (average over 50 random runs)

Figure 5: Analysis the SCP on multi-concepts personalization. The embedding of V^*_{man} **is fixed** while the embedding of V^*_{dog} **is varied** across different training steps. Prompt: 'A photo of a V^*_{man} touching a V^*_{dog} beside a park bench'. See Figure 19 and 20 for more examples.

It can be observed from Figure 5a that at early steps (e.g., < 400), V^*_{man} tends to dominate the prompt, resulting in images that primarily capture the 'man' concept, consistent with the SCP observed in single-concept settings. However, as the training progresses and $V^*_{dog,step}$ captures more distinctive features of the personalized dog concept, it begins to overshadow V^*_{man}, leading to outputs that predominantly depict only the dog, effectively suppressing the presence of the other concept. As shown in Figure 5b, the alignment score with anchor man image drops gradually while that with anchor dog image increases over time further confirming the dominance of V^*_{dog} over V^*_{man}.

This trend, observed consistently across multiple settings as shown in Figure 19 and 20, highlights the intricate nature of SCP in multi-concept personalization. It suggests that SCP not only persists but can intensify when multiple personalized concepts are involved, presenting a challenging but potentially fruitful direction for future research.

4 Training-Free Embedding Adjustment for SCP

As demonstrated in the previous section, SCP exhibits two-way impacts that vary depending on the nature of the context prompt. This variability makes it challenging to devise a universal solution that mitigates the negative effects of SCP while preserving its beneficial aspects across all inference prompts. Recall that in the earlier analysis, we observed that the learned embedding V^* often drifts from its original semantic anchor c due to unconstrained optimisation—resulting in significant shifts in both magnitude and direction. This raises a natural question: *Can we reverse this semantic shift at test time by adjusting V^*, without modifying the personalization method?*

A key advantage of this approach is that it is training-free and can be applied post hoc to any pre-trained personalization method, regardless of whether it is based on Textual Inversion or DreamBooth. Surprisingly, this simple adjustment proves to be highly effective.

Embedding Adjustment Given a pre-trained embedding matrix M that includes a learned token V^* (as in Textual Inversion), and a target concept c toward which we wish to regularise, we propose to adjust M_{V^*} by aligning both its magnitude and direction with M_c. This is achieved by first normalising the vectors and then applying Spherical Linear Interpolation (SLERP) [27] to interpolate the direction of M_{V^*} towards M_c, which is known to be effective in high-dimensional vector spaces.

$$\hat{M}_{V^*} = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)}\tilde{M}_{V^*} + \frac{\sin(\alpha\theta)}{\sin(\theta)}\tilde{M}_c \tag{1}$$

Here, θ is the angle between the normalized vectors $\tilde{M}_c$ and $\tilde{M}_{V^*}$, and $\alpha \in [0, 1]$ controls the rotation factor, where the bigger α is, the more the embedding is rotated towards M_c. The normalisation vectors are defined as $\tilde{M}_{V^*} = \beta \|M_c\| \frac{M_{V^*}}{\|M_{V^*}\|}$ and $\tilde{M}_c = \beta \|M_c\| \frac{M_c}{\|M_c\|}$ where β is the scaling factor to control the magnitude of the embedding relative to the reference concept c.

In Dreambooth-based personalization, because the embedding matrix M is not updated during the optimisation, we propose to adjust at the prompt level instead of the token level. More specifically,

given a prompt $\lfloor p, V^* \rfloor$ and a target prompt $\lfloor p, c \rfloor$, we obtain the two embeddings $\tau(\lfloor p, V^* \rfloor)$ and $\tau(\lfloor p, c \rfloor)$ from the text encoder τ_ϕ and then adjust the embedding of $\lfloor p, V^* \rfloor$ by using the above equation on every token in the prompt.

$$\hat{\tau}(\lfloor p, V^* \rfloor)[i] = \frac{\sin((1-\alpha)\theta_i)}{\sin(\theta_i)} \tilde{\tau}(\lfloor p, V^* \rfloor)[i] + \frac{\sin(\alpha\theta_i)}{\sin(\theta_i)} \tilde{\tau}(\lfloor p, c \rfloor)[i] \tag{2}$$

where i indexes each token in the prompt, and θ_i is the angle between the i-th token embeddings of the two prompts after normalisation. This method enables a test-time adjustment of semantic drift without retraining, making it a lightweight and broadly applicable solution to mitigating SCP effects.

5 Experiments

5.1 Experimental Setup

Reference Images. We use a subset of 9 concepts from the CustomConcept101 (CC101) dataset as in the original paper [16], each of which has 3-15 images, including 'Barn', 'Tortoise plushy', 'Teddy-Bear', 'Wooden Pot', 'Dog', 'Cat', 'Flower', 'Table', 'Chair' subjects. For the human concept, we use a subset of 10 concepts from the CelebA-HQ dataset [19], which includes 10 identities with 10-15 images per subject. Sample images from the two datasets are shown in Figures 8 and 9.

Prompts. We collect complex prompts from the CC101 dataset, where each prompt contains at least two concepts, e.g., 'a watercolor painting of V^* tortoise plushy on a mountain'. For the human concept, we create a list of 17 prompts, where each prompt contains the main concept and a complex context/action, e.g., 'A photo of a V^* wearing glasses and writing in a red notebook'. The prompts can be found in Table 4.

Table 1: Analysis of the effect of rotation factor.

Method	CLIP_T^c	CLIP_T^p	CLIP_T^f	CLIP-I
TI	17.4 ± 1.8	21.4 ± 1.1	26.5 ± 1.8	73.5 ± 4.7
$\alpha = 0.20$	17.8 ± 1.9	21.7 ± 1.1	27.5 ± 1.9	71.1 ± 5.7
$\alpha = 0.25$	18.0 ± 1.9	21.8 ± 1.0	27.8 ± 1.9	69.6 ± 5.7
$\alpha = 0.30$	18.0 ± 2.0	21.9 ± 1.1	28.0 ± 1.9	67.3 ± 5.8
$\alpha = 0.35$	18.2 ± 2.0	21.9 ± 1.1	28.2 ± 1.8	63.6 ± 5.3

Metrics. In addition to the CLIP-T and CLIP-I alignment scores introduced in Section 3.2, we also use the DINO **image-image** alignment score [3] to evaluate the alignment between the generated images and the reference images.

5.2 The Effect of Hyper-parameters

An important question is how to choose the hyper-parameters α and β appropriately, or whether they should be adapted based on the input prompt. Interestingly, our experiments reveal a clear pattern in the performance of the proposed method across a range of α and β values, providing practical guidance for their selection. For simplicity and consistency, we set $\alpha = 0.2$ and $\beta = 1.5$ as default values, which consistently deliver robust performance across diverse prompts.

Table 2: Improvement of our TEA over its baselines counterpart.

Method	CLIP_T^p	CLIP_T^f	CLIP-I	DINO-I
CC101				
TI+TEA	0.55	0.64	-2.34	-3.81
DB+TEA	0.77	1.87	4.57	0.59
CD+TEA	-0.13	0.34	0.37	0.24
CelebA				
TI+TEA	0.33	0.57	-2.41	-1.78
DB+TEA	0.51	0.64	-2.37	-2.27
CD+TEA	-0.12	0.09	1.56	3.09

Figure 6 shows the performance of our adjustment method over a range of β values with fixed $\alpha = 1.0$ (no rotation). It can be seen that when β is too small (i.e., $\beta < 0.5$, meaning $\|\hat{M}_{V^*}\| < 0.5\,\|M_c\|$) or too large (i.e., $\beta > 5.0$, meaning $\|\hat{M}_{V^*}\| > 5.0\,\|M_c\|$), the generated images become less aligned with the ground truth indicating by the significant drop in $S_I(\hat{x}, x_{gt})$, suggesting that V^* has lost its personalized information. However, the alignment $S_I(\hat{x}, x_{gt})$ is relatively stable when β is in the range of $[1.0, 5.0]$, suggesting that the personalized concept can be effectively captured without extreme scaling V^*. As a practical choice, we simply set $\beta = 1.5 \approx \frac{\|M_{V^*}\| + \|M_c\|}{2\|M_c\|}$ (which is a middle value interpolated from $\|M_c\|$ to $\|M_{V^*}\|$), as the default setting.

Table 1 shows the performance over a range of α values with fixed $\beta = 1.5$. Unlike the scaling parameter, the rotation factor α is more sensitive and significantly impacts prompt alignment capability. Increasing α generally improves the model's ability to capture context, as reflected in the CLIP_T^f

score, which increases from 26.5 to 28.2 (an improvement of 1.7), and the CLIP_T^p score, which rises from 21.4 to 21.9 (a gain of 0.5). While this comes at a minor cost to visual fidelity, with the CLIP-I score dropping from 73.5 to 71.1 when $\alpha = 0.2$, the generated images still maintain high visual quality, as shown in Figures 16 and 17.

These findings highlight the critical role of the rotation factor α, which directly controls the semantic alignment between V^* and the target concept c. Higher α values encourage better prompt alignment by rotating V^* closer to c, while the scaling factor β should remain within a moderate range to prevent excessive distortion of the learned visual concept.

5.3 Evaluation Results

We conducted an extensive evaluation to validate the effectiveness of the proposed method (Test-time Embedding Adjustment, TEA) when combined with pre-trained personalization methods such as Textual Inversion (TI) [9], DreamBooth (DB) [25], and CustomDiffusion (CD) [16].

The results in Table 6 show that all baselines 'enjoy' a significant improvement on prompt alignment when combined with our 'TEA', with the average CLIP_T^f score increase of 0.54 for TI with TEA, 0.33 for CD with TEA, and 1.9 for DB with TEA, demonstrating substantial gains in text-image alignment. Notably, this improvement is consistent across most concepts as shown in Figure 7, indicating that our method is robust to diverse personalization scenarios, even with fixed, easily chosen hyper-parameters. In terms of visual quality, our method also shows a noticeable boost, with a 4.6 improvement in the CLIP-I score from DB to DB+EA or 0.36 improvement for CD+TEA on the CC101 dataset. Our method also shows interesting effects over the baseline counterparts, as shown in Figure 15, where the SCP in DB lead to the distorted generated images. Our method successfully corrects the semantic embedding and generates more coherent and realistic images. In the CelebA dataset, our method still shows its effectiveness in improving prompt alignment, with a 0.64 improvement in the CLIP_T^f score for DB+TEA. There is also a trade-off in image alignment with a 2.37 drop in the CLIP-I score, however, the generated images still maintain high visual quality, as shown in Figures 16 and 17.

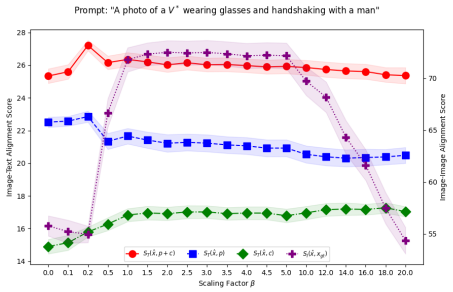


Figure 6: Analysis of the effect of scaling factor.

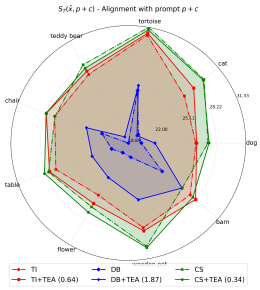


Figure 7: Comparison on prompt alignment of our TEA over its baselines counterpart on CC101 dataset. Refer to Table 6 for detailed numbers.

6 Conclusion

In this paper, we identified the *Semantic Collapsing Problem* (SCP) in generative personalization, where personalized tokens lose their original semantic meaning and dominate other concepts in complex prompts. We traced this issue to unconstrained optimisation, which allows the learned token embedding to drift in direction and magnitude, disrupting prompt interpretation.

To address this, we proposed a training-free test-time embedding adjustment that realigns the personalized embedding with its original semantic context, significantly improving text-image alignment without modifying model weights. Our method, compatible with diverse personalization approaches like Textual Inversion and DreamBooth, demonstrates substantial gains in prompt consistency and image fidelity across varied use cases.

Beyond introducing the SCP problem and providing a practical, straightforward solution, this work also lays the foundation for future research. Potential directions include developing more adaptive approaches that dynamically adjust embeddings based on prompt complexity or context, as well as integrating additional constraints during personalization to better control embedding updates.

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–12, 2023.

[2] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

[4] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8619–8628, 2024.

[5] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023.

[6] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793*, 2023.

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

[8] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, and Zhendong Mao. Dreamidentity: enhanced editability for efficient face-identity preserved image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1281–1289, 2024.

[9] 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.

[10] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7323–7334, 2023.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] Hugging Face. Dreambooth training example. `https://github.com/huggingface/diffusers/blob/main/examples/dreambooth/train_dreambooth_lora.py`. dev 0.33.0.

[13] Zeyinzi Jiang, Chaojie Mao, Yulin Pan, Zhen Han, and Jingfeng Zhang. Scedit: Efficient and controllable image diffusion generation via skip connection editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pp. 8995–9004, 2024.

[14] Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Alexander Teare. An image is worth multiple words: Discovering object level concepts using multi-concept prompt learning. In *Forty-first International Conference on Machine Learning*, 2024.

[15] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *European Conference on Computer Vision*, pp. 253–270. Springer, 2024.

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1931–1941, 2023.

[17] Yuheng Li, Haotian Liu, Yangming Wen, and Yong Jae Lee. Generate anything anywhere in any scene. *arXiv preprint arXiv:2306.17154*, 2023.

[18] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter. *arXiv preprint arXiv:2312.00330*, 2023.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[20] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

[21] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in Neural Information Processing Systems*, 36:79320–79362, 2023.

[22] 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. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[25] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

[26] Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, and Ali Mahdavi-Amiri. Clic: Concept learning in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6924–6933, 2024.

[27] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pp. 245–254, 1985.

[28] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

[29] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[30] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

[31] Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–10, 2023.

[32] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024.

[33] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pp. 1–20, 2024.

[34] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7754–7765, 2023.

[35] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8682–8692, 2024.

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

[37] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. Multibooth: Towards generating all your concepts in an image from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10923–10931, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a separate "Limitations and Future Work" section in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed instructions to reproduce the experimental results in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the data and code in the anonymous repository at `https://anonymous.4open.science/r/Embedding-Adjustment`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the experimental setting in Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For each prompt, we generate 50 to 100 images to avoid the randomness of the generation process and report the mean (and standard deviation) of the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the computational settings in Appendix E.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [NA]

Justification: We do not aware of any ethical issues in the paper.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper aims to improve personalization methods for text-to-image generation. We do not aware of any potential negative societal impacts beyond the general risks already associated with recent text-to-image generation models.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper aims to improve personalization methods for text-to-image generation. We do not aware of any potential negative societal impacts beyond the general risks already associated with recent text-to-image generation models. Therefore, we did not describe any safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We did properly cite the original papers for the code and datasets used in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We create a new list of prompts to generate images in the paper and provide detailed prompts in the appendix.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We utilized LLMs to suggest the sample prompts and provided the prompts in the appendix.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

Table of Contents

A Related Work **20**
A.1 Diffusion Models 20
A.2 Generative Personalization. 20
A.3 Challenges in Generative Personalization 21

B Further Quantitative Results **22**

C Qualitative Results **22**

D Limitations and Future Work **22**

E Experimental Setting **23**
E.1 Dataset Construction 23
E.2 Evaluation Metrics 24
E.3 Computational Settings 28

A Related Work

A.1 Diffusion Models

Given a text-to-image diffusion model ϵ_θ, where $\epsilon_\theta(x_t, t, p)$ represents the predicted noise at time step t given the textual embedding $\tau(p)$ of a prompt p and the noisy intermediate vector x_t [11, 29, 24], the model is trained by minimizing the following objective:

$$\mathcal{L} = \mathbb{E}_{(x,p)\sim p_{\text{data}}, t\sim\mathcal{U}[0,T], \epsilon\sim\mathcal{N}(0,\mathbf{I})} \left[\left\| \epsilon - \epsilon_\theta(\alpha_t x + \sqrt{1-\alpha_t}\epsilon, t, p) \right\|_2^2 \right] \tag{3}$$

Here, x and p denote the input image and its associated prompt, respectively, while ϵ is the Gaussian noise sampled from a standard normal distribution. The intermediate input $x_{t,\epsilon} = \alpha_t x + \sqrt{1-\alpha_t}\epsilon$ is obtained from the forward diffusion process. For simplicity, we use the notation $\mathbb{E}_{x,p,t,\epsilon}[\cdot]$ to represent the expectation over the input data x, the prompt p, the diffusion time step t, and noise ϵ.

A.2 Generative Personalization.

Generative personalization task aims to capture and utilise personal concepts which are implicitly shared in a reference set of images as generative conditions, which are not easily expressible in the input prompt. There are several categories of generative personalization [2] classification based on the type of generative conditions, such as subject-driven [9, 25, 7, 16, 32], person-driven [33, 31, 8, 6], style-driven [28, 18, 4] or image-driven [23, 34, 35]. Personalizing a T2I diffusion model from only a few examples presents several well-known challenges, including language drift, limited expressiveness of generative conditions, entanglement of concepts, and conditional misalignment.

Textual Inversion and Dreambooth. While there are many personalization methods have been proposed, they can be traced back to the two representative methods: Textual Inversion (TI) [9] and DreamBooth [25]. Mathematically, given a set of personal images $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ and a pre-trained T2I model ϵ_θ, the goal is to identify a textual embedding v^* associated with a specific verbalizable keyword V^* (e.g., 'sks', '<new>', etc.). This keyword represents the implicit visual concept shared in the reference set $\mathcal{X}$, enabling the model to generate images with the personal concept using any textual prompt p containing the keyword V^*, e.g., $\lfloor p, V^* \rfloor$ = 'A photo of V^* playing on a beach', where $\lfloor ., . \rfloor$ is the sentence construction operator.

Textual Inversion (TI) [9] is a pioneering method that proposes obtaining a textual embedding by minimising the following objective:

$$\min_{v^*} \mathbb{E}_{x,p,\epsilon,t}\left[\|\epsilon - \epsilon_\theta\left(x_{t,\epsilon}, t, \lfloor p, V^* \rfloor\right)\|_2^2\right] \tag{4}$$

Here, $p \sim \mathcal{T}$ is a template prompt sampled from a set of predefined neutral prompts $\mathcal{T}$, such as {'a photo of a', 'a high-quality photo of a', $\cdots$}. In TI, only the embedding v^* is learned, while all other parameters, such as the model ϵ_θ or textual encoder τ, remain fixed. Although this method is parameter-efficient, the learned embedding may not be sufficiently representative to capture the true visual concept in the reference set. Building on TI, Dreambooth [25] suggests fine-tuning not only the embedding v^* (i.e., by fine-tuning the textual encoder τ) but also the model parameters θ by minimizing the following objective:

$$\min_{\theta,v^*} \mathbb{E}_{x,p,x_{pr},p_{pr},\epsilon,\epsilon',t}\left[\|\epsilon - \epsilon_\theta\left(x_{t,\epsilon}, t, \lfloor p, V^* \rfloor\right)\|_2^2 + \lambda \left\|\epsilon' - \epsilon_\theta\left(x^{pr}_{t',\epsilon'}, t', p^{pr}\right)\right\|_2^2\right] \tag{5}$$

In this context, x^{pr} and p^{pr} are the prior-preservation image and its associated prompt, respectively, which help prevent the model from overfitting to the small reference set with a tradeoff hyper-parameter λ.

A.3 Challenges in Generative Personalization

Language Drift. Language drift or overfitting occurs due to the limited number of reference images and results in the incorporation of irrelevant elements and the neglect of the textual context within the outputs. Prior works address this issue by introducing preservation mechanisms such as prior-preservation loss [25], locking concept-specific parameters [30] and regularising model weights [10, 21].

Limited Expressiveness of Generative Conditions. This occurs due to the limited expressiveness of the original textual format and the limited number of tokens allowed in each condition. A common mitigation approach is to use multi-modal conditions such as image-image and sketch-image. To enable pre-trained T2I models to accept new types of conditions and generate in conjunction with the current text prompt conditioning, previous works have attempted to incorporate an additional encoder [36], or add a new adapter module to align internal knowledge of the model with the new condition [20, 13].

Entanglement of Concepts. The reference image set might include samples that contain both the intended concept and other irrelevant concepts. To effectively isolate and extract the intended concept from the reference set, previous works have employed explicit masks [1, 14, 26] and additional data with the personalized concept [17]. Alternatively, Disenbooth [5] proposed to mitigate the influence of background elements in the reference set by disentangling the identity and background of the reference set.

Conditional Misalignment. Conditional misalignment is a long-standing problem in the field of conditional generation and is a consequence of the above mentioned challenges, with its cause traced back to the limited alignment power of the original generative model. A largely unaddressed gap in prior personalization work is the potential drift or misalignment of the textual embedding itself during concept learning. We refer this phenomenon as *semantic collapse* where the learned concept token is still faithful to the visual reference, but fails to retain any meaningful textual semantics and eventually collapses to a simplified form. Existing methods focus on controlling how the model changes through regularisation, but not explicitly ensure that the learned personalized token embedding remains semantically aligned with the context. In this work, we tackle the issue with a lightweight training-free test-time embedding adjustment which is broadly applicable to different personalized methods.

B Further Quantitative Results

In this section, we provide further quantitative results on the CustomConcept101 dataset and the CelebA dataset. More specifically, we provide the additional results as follows:

- Additional results showing the semantic collapsing problem in the textual domain measured by different distances in Figure 10.
- Additional results showing the semantic collapsing problem in images generated by Textual Inversion and DreamBooth in Figures 11 and 12 respectively.
- Empirical evidence of the semantic shift of the embedding of the entire prompt on DreamBooth in Figure 14.
- Detailed alignment scores of our TEA over its baselines counterpart on the CS101 and CelebA datasets in Table 6 and Table 7 respectively.

C Qualitative Results

We provide the additional qualitative results as follows:

- The cool effects of our method over the baseline counterpart in Figure 15, where the SCP in DB lead to the distorted generated images. Our method successfully corrects the semantic embedding and generates more coherent and realistic images.
- Qualitative results of the proposed method over its baselines counterpart on the CS101 and CelebA datasets in Figures 17 and 16 respectively.
- Illustration of the SCP on multiple concepts setting in Figures 19 and 20. The alignment scores are shown in Figure 18.

D Limitations and Future Work

Limitations on Evaluation. While we tried our best to design a comprehensive set of experiments to evaluate the proposed method and baselines, there are still some limitations that we would like to address in the future. First, the model used in this paper is the Stable Diffusion v1.5 model, which is one of the most popular pre-trained models in this field. While the proposed method is general and can be applied to other models, we did not test it on larger models, e.g., SDXL or SD version 2.0. Second, we chose three baselines, including Textual Inversion [9], DreamBooth [25], and Custom Diffusion [16], which are among the most popular personalization methods in the field. While the proposed method is general and can be applied to other personalization methods, it might be interesting to apply it to other approaches as well.

Limitations on Methodology. We believe that the insights and understanding provided by this paper, especially the analysis of the Semantic Collapsing Problem, are the most important contributions. Based on this analysis, we propose a simple method to adjust the embedding vectors at test time. While this method has clear advantages, such as simplicity, generalizability, and no additional training requirement, the simplicity of the method itself might be a limitation. There are still trade-offs between alignment with the input prompt and alignment with the visual concept, which depend on the hyper-parameters, suggesting that fixed hyper-parameters might not be optimal for all prompts.

Future Work. We believe that the insights and understanding provided by this paper, especially the analysis of the Semantic Collapsing Problem, can guide future research on this topic. For example, integrating additional constraints to restrict semantic shift during the fine-tuning phase, rather than relying solely on test-time adjustment, is a promising direction. This approach could directly produce adjusted and bounded embedding vectors that retain the original semantic meaning of the base concept.

In this work, we provide a simple method to adjust the embedding vectors at test time. This adjustment is applied equally across all dimensions of the embedding vectors. However, we believe that this is not the optimal way to adjust embedding vectors, as each dimension of the embedding vector

has different meanings and importance. Therefore, a more sophisticated method that considers the importance of each dimension could also be a promising direction.

In Section 5.2, we provide an analysis of the impact of hyper-parameters on the performance of the proposed method. It has been shown that the performance of the proposed method is sensitive to the rotation factor α: the larger the α, the more the embedding vector is rotated toward the target concept, improving alignment with the generated images but potentially reducing alignment with the visual concept. In this work, we simply use the same hyper-parameters for all settings. We believe that a search algorithm could be applied to find the optimal hyper-parameters for each prompt, with a stop condition based on the desired alignment with the input prompt.

E Experimental Setting

E.1 Dataset Construction

Contextual Prompts for Measuring Semantic Collapsing. Recall our hypothesis: a personalized keyword V^*, initialized from a base concept c to capture a specific visual target v_{gt}, tends to lose its original semantic meaning and dominate arbitrary contexts p when used in complex prompts. This phenomenon, which we refer to as *semantic shift*, can be directly assessed by comparing the embedding vectors M_{V^*} and M_c. However, this is challenging due to the use of contextualized text embeddings in modern LLMs and diffusion models, where the surrounding context significantly shapes the final representation of each token.

To address this, we propose evaluating the semantic shift of V^* relative to c in the presence of a diverse set of contextual prompts. Specifically, we define a prompt set $A = \{a_1, a_2, \cdots, a_n\}$, constructed by querying a large language model (LLM) with the following instruction:

> `Write 200 sentences with diverse topics and contents. Each sentence should be 10-30 words long and must include the keyword` c`.`

This approach allows us to measure how well the learned embedding M_{V^*} retains the original semantic characteristics of c across varied contexts, providing a robust test for the semantic collapsing problem.

To further examine the dominance effect of V^*, we introduce a complementary set of simple prompts, denoted as $P^{\text{simple}}_{V^*}$. This set consists of 200 straightforward sentences where V^* is the clear focal point, such as "`a photo of a` V^*", "`a portrait of a` V^*", etc. These simple prompts serve as a baseline for assessing the degree to which V^* overshadows other contextual elements in the generated outputs.

Sample sentences from both prompt sets are provided in Table 3, and the full dataset, along with all prompt templates, is available in the anonymous repository at `https://anonymous.4open.science/r/Embedding-Adjustment`.

Dataset for Evaluating Personalization Performance. We use a subset of 9 concepts from the CustomConcept101 dataset as in the original paper [16], each of which has 3-15 images, including 'Barn', 'Tortoise plushy', 'Teddy-Bear', 'Wooden Pot', 'Dog', 'Cat', 'Flower', 'Table', 'Chair' subjects. For the human concept, we use a subset of 10 concepts from the CelebA-HQ dataset [19], which includes 10 identities with 10-15 images per subject. Sample images from the CustomConcept101 and CelebA-HQ datasets are shown in Figure 8 and Figure 9, respectively.

To assess complex prompt handling, we compile a set of multi-concept prompts from the CustomConcept101 dataset. Each prompt is designed to include two to three distinct elements, encouraging the model to balance multiple visual contexts. For instance:

- "V^* tortoise plushy sitting at the beach with a view of the sea"
- "a watercolor painting of V^* tortoise plushy on a mountain"

These prompts contain a primary subject V^* and one or more contextual elements (context p), allowing us to measure the model's ability to preserve the personalized concept while maintaining accurate context alignment.

Table 3: Sample sentences for P_{V^*} and $P_{V^*}^{\text{simple}}$. Set P_c can be constructed by replacing the word c with V^* in the prompt of P. All the data and prompts can be found in the anonymous repository.

Set	Sample Sentences
P_{V^*}	'A V^* walked his dog through the park every morning before sunrise' 'Despite the heavy rain, a V^* stood patiently waiting for the bus' 'In the small village, a V^* known for his kindness helped everyone' 'After twenty years of dedicated service, a V^* retired from his factory job' 'While climbing Mount Everest, a V^* discovered the true meaning of perseverance' 'During the concert, a V^* in the front row sang along to every song' 'At the crowded marketplace, a V^* sold handcrafted jewelry made from local materials' 'Throughout history, a V^* with vision has often changed the course of events' 'Behind every successful company, there is often a V^* with an innovative idea' 'Within the ancient temple, a V^* prayed silently for his family's wellbeing'
$P_{V^*}^{\text{simple}}$	'A photo of a V^*' 'A rendering of a V^*' 'A cropped photo of a V^*' 'A portrait of a V^*' 'A close-up shot of a V^*' 'A full-body image of a V^*' 'A black-and-white photograph of a V^*' 'A candid shot of a V^*' 'A digital illustration of a V^*' 'A stylized caricature of a V^*'

Sample prompts are provided in Table 4, and the full dataset, along with all prompt templates, is available in the anonymous repository at `https://anonymous.4open.science/r/Embedding-Adjustment`.

E.2 Evaluation Metrics

Personalization Metrics. We use the CLIP-T **text-image** alignment score [22] to evaluate the alignment between the generated images $\hat{x}$ and the prompts. To have a better understanding of which part of the prompt contributes to construct the generated images, we break down each prompt into multiple segments/concepts and calculate the alignment score for each segment/concept, i.e., 'CLIP_T^f'/'CLIP_T^c'/'CLIP_T^p' denotes the alignment score of the full prompt, the first segment—the personal concept, and the second segment—the context, respectively. We use the CLIP-I **image-image** alignment score [22] and DINO **image-image** alignment score [3] to evaluate the alignment between the generated images and the reference images.

Semantic Shifting Metrics. Given the two sets P and Q (i.e., $P = P_{V^*} = \{\tau(\lfloor a_i, V^* \rfloor)\}$ and $Q = P_c = \{\tau(\lfloor a_i, c \rfloor)\}$ which are the embeddings of the prompts in the prompt sets P_{V^*} and P_c respectively), we propose to use the following metrics to measure the difference between the two sets.

L2 distance measures the Euclidean distance between two points in a vector space.

$$d_{L2}(p_i, q_i) = \|p_i - q_i\|_2 \tag{6}$$

Hausdorff distance measures the maximum distance from any point in P to the nearest point in Q. In other words, it measures the greatest distance from a point in one set to the nearest point in another set.

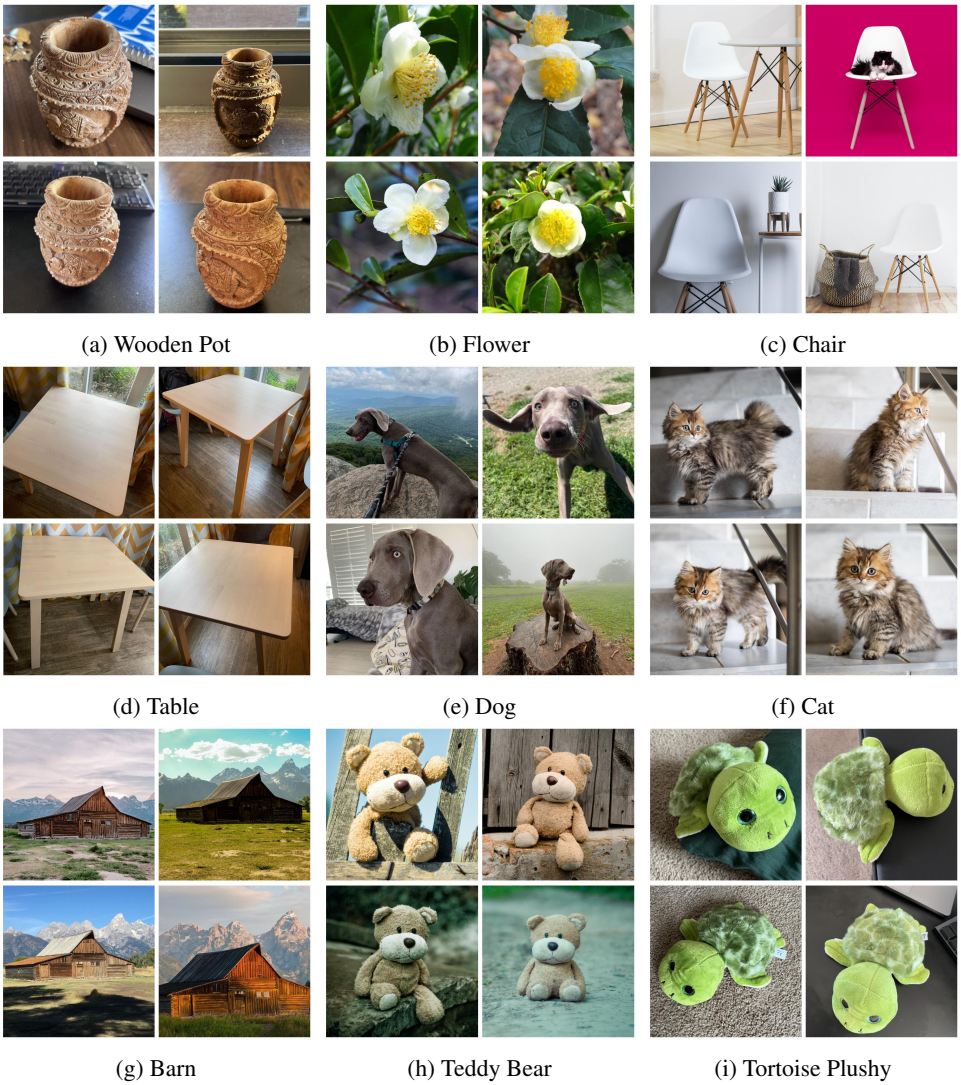
$$d_H(P, Q) = \max(\max_{p_i \in P} \min_{q_j \in Q} d_{L2}(p_i, q_j), \max_{q_i \in Q} \min_{p_j \in P} d_{L2}(q_i, p_j)) \tag{7}$$

Table 4: Sample prompts to generate personalized images. Each prompt consists of a main subject V^* and a context p. All the data and prompts can be found in the anonymous repository.

Set	Sample Sentences
CelebA	'A photo of a V^* wearing glasses and handshaking with a man' 'A photo of a V^* wearing glasses and handshaking with a woman' 'A photo of a V^* wearing glasses and handshaking with an old man' 'A photo of a V^* wearing glasses and handshaking with a kid' 'A photo of a V^* wearing glasses and holding a dog' 'A photo of a V^* wearing glasses and holding a cat' 'A photo of a V^* wearing glasses and holding a red book' 'A photo of a V^* wearing glasses and holding a red phone' 'A photo of a V^* wearing glasses and sitting on a red chair' 'A photo of a V^* wearing glasses and lying on a red bed' 'A photo of a V^* wearing glasses and writing in a red notebook' 'A photo of a V^* wearing glasses and drinking a Coco Cola can' 'A photo of a V^* wearing glasses and lifting weights' 'A photo of a V^* wearing glasses and cycling' 'A photo of a V^* wearing glasses and kicking a football' 'A photo of a V^* wearing glasses and playing a guitar' 'A photo of a V^* wearing glasses and eating a pizza'
CustomConcept101	'V^* in snowy ice' 'V^* in blooming sunflower field' 'V^* on a boat in the sea' 'V^* on top of a mountain' 'V^* made of crochet' 'V^* in a garden' 'a floor lamp on the side of V^* ' 'V^* and a table with chocolate cake on it' 'a puppy sitting on a V^* ' 'a cat sitting on a V^* ' 'a squirrel sitting on a V^* ' 'a deer grazing near a V^* ' 'a teddy bear on a V^* ' 'a photo of a V^* in Van Gogh style'
Multi-Concept	'A photo of a V^*_{man} wearing glasses and kissing a V^*_{dog} ' 'A photo of a V^*_{man} wearing glasses and handshaking with a V^*_{dog} ' 'A photo of a V^*_{man} wearing a hat and hugging a V^*_{dog} ' 'A photo of a V^*_{man} walking a V^*_{dog} on a road with a car behind' 'A photo of a V^*_{man} holding a V^*_{dog} beside a car' 'A photo of a V^*_{man} touching a V^*_{dog} beside a park bench' 'A photo of a V^*_{man} feeding a V^*_{dog} with a bowl of flowers'

Table 5: Inter-set and intra-set distances for different distance metrics

Distance	**Inter-set Distance** $d(P,Q)$	**Intra-set Distance** $d(P,P)$
L2	$\frac{1}{n}\sum_{i=1}^{n}\|p_i-q_i\|_2^2$	$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|p_i-p_j\|_2^2$
Hausdorff	$\max(\max_{p_i\in P}\min_{q_j\in Q} d_{L2}(p_i,q_j),$ $\max_{q_i\in Q}\min_{p_j\in P} d_{L2}(q_i,p_j))$	$\max_{p_i\in P}\min_{p_j\in P, p_j\neq p_i} d_{L2}(p_i,p_j)$
Mahalanobis	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(p_i-\mu_P)^T\Sigma_P^{-1}(p_i-\mu_P)}$	$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{(p_i-p_j)^T\Sigma_P^{-1}(p_i-p_j)}$
KL	$D_{KL}(P \mid\mid Q)$	$D_{KL}(P \mid\mid P)$



(a) Wooden Pot (b) Flower (c) Chair

(d) Table (e) Dog (f) Cat

(g) Barn (h) Teddy Bear (i) Tortoise Plushy

Figure 8: Sample images from the CustomConcept101 dataset.

Mahalanobis distance measures how far the point p_i is from the center of the set P, taking into account the correlation between the dimensions of the set. Unlike the L2 distance which treats all dimensions equally, the Mahalanobis distance adapts to the shape and spread of the set P.

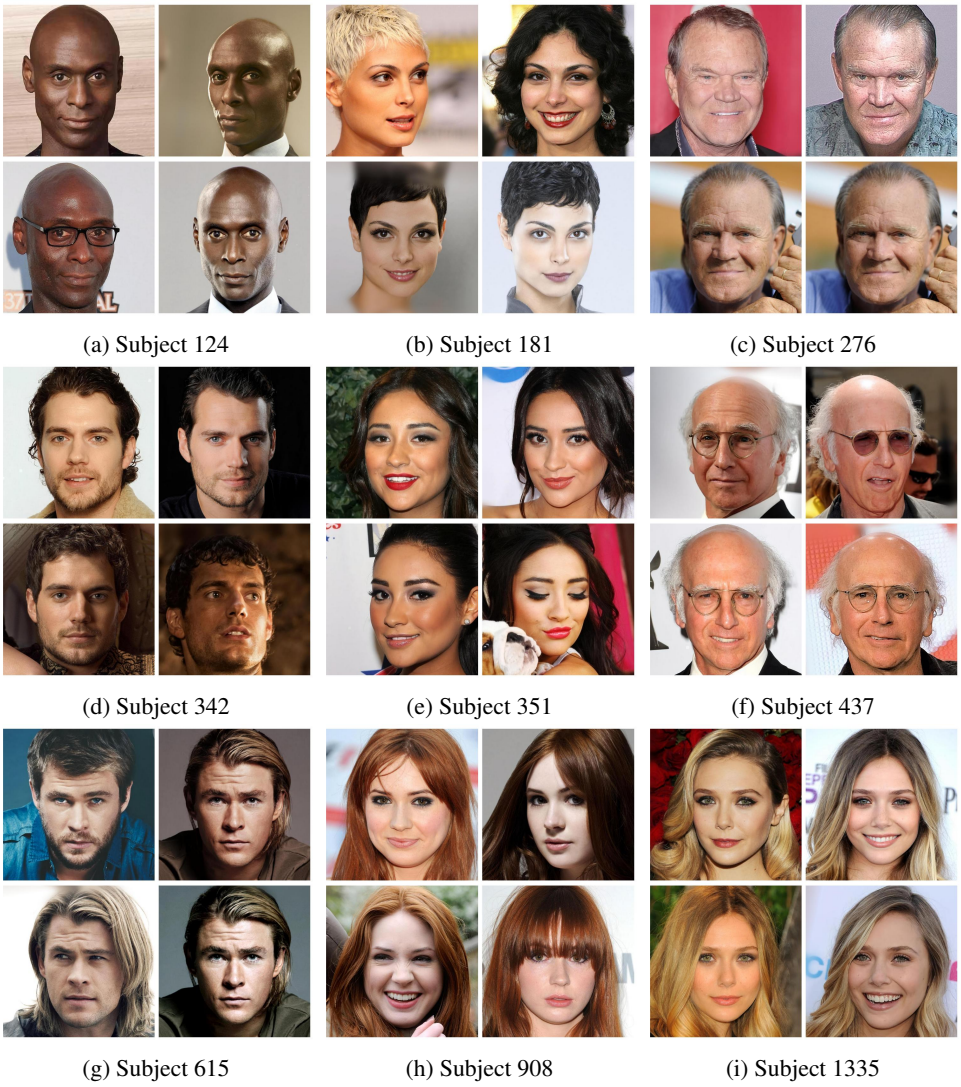
$$d_M(p_i, P) = \sqrt{(p_i - \mu_P)^T \Sigma_P^{-1} (p_i - \mu_P)} \tag{8}$$

where μ_P is the mean of the set P and Σ_P is the covariance matrix of the set P.

KL divergence . We propose to measure the relative relationship between each data point to the entire set by using the Normalized Temperature-scaled Softmax function

$$p(p_i \mid p_j, P) = \frac{\exp(\text{sim}(p_i, p_j)/T)}{\sum_{p_k \in P} \exp(\text{sim}(p_i, p_k)/T)} \tag{9}$$

where T is the temperature parameter. $p(p_i \mid p_j, P)$ measures the the relative relationship between p_j and the anchor p_i in comparison to the entire set P. From that, we can have $p(p_i \mid P) = \{p(p_i \mid p_j, P)\ \forall p_j \in P\}$ to represent the relative relationship between p_i and the entire set P. Similarly,



(a) Subject 124 (b) Subject 181 (c) Subject 276

(d) Subject 342 (e) Subject 351 (f) Subject 437

(g) Subject 615 (h) Subject 908 (i) Subject 1335

Figure 9: Sample images from the CelebA dataset.

$p(q_i \mid Q) = \{p(q_i \mid q_j, Q)\ \forall q_j \in Q\}$ to represent the relative relationship between q_i and the entire set Q.

The KL divergence between P and Q is then defined as:

$$D_{KL}(P \mid\mid Q) = \sum_{i=1}^{n} p(p_i \mid P) \log \frac{p(p_i \mid P)}{p(q_i \mid Q)} \tag{10}$$

which measures the difference between the two distributions P and Q. The higher the KL divergence, the more different the two distributions are, the more semantic shifting the learned embedding v^* has.

Alignment Metrics. The primary objective of generative personalization is to produce visually compelling images that accurately capture the unique characteristics of a personalized concept from a reference set, while maintaining semantic alignment with the input textual prompt.

To evaluate this, we use two key metrics based on the CLIP model [22]:

Visual Fidelity: We measure the alignment between the generated image and a ground-truth image using the CLIP image-image alignment score. A higher score indicates a closer match to the reference, reflecting better preservation of the personalized visual features.

Prompt Consistency: We assess the alignment between the generated image and the input textual prompt using the CLIP text-image alignment score. A higher score indicates that the generated image more accurately reflects the intended context and details of the input text.

However, complex prompts often contain multiple concepts, making a single text-image alignment score insufficient to capture the nuanced relationship between the personalized concept V^* and its broader context p. To address this, we separately compute alignment scores for:

Main Concept Alignment (V^*): Measuring the fidelity of the personalized concept itself.

Context Alignment (p): Evaluating how well the broader contextual elements are represented.

This multi-level evaluation provides a more comprehensive understanding of how well the generated images capture both the personalized visual identity and the intended scene context.

For all evaluations, we use the implementation provided in the TorchMetrics library, available at `https://lightning.ai/docs/torchmetrics/stable/multimodal/clip_score.html`.

E.3 Computational Settings

All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24GB of memory, using the Stable Diffusion v1.5 model as the base model. To prevent memory overflow, we fine-tune the model with Textual Inversion (TI), DreamBooth (DB), and Custom Diffusion (CD) using a batch size of 1 across all methods.

For the DreamBooth LoRA method, we follow the recommended settings from the Diffusers' example page, using a learning rate of 1e-4, rank 4, and enabling text encoder training for improved performance. Textual Inversion is fine-tuned with a learning rate of 5e-4, while Custom Diffusion uses a more conservative learning rate of 5e-6.

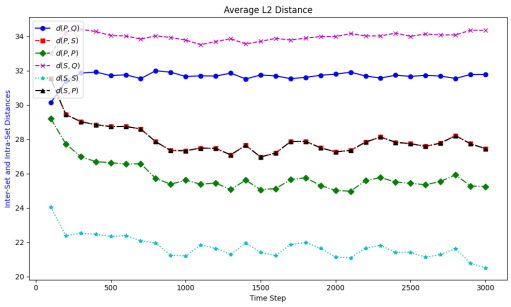
All code implementations are adapted from the Hugging Face Diffusers library.

Table 6: Results on CustomConcept101 dataset, tort* means tortoise plushy, teddy* means plushy teddy bear, wpot* means wooden pot. The first/second metric is the CLIP_T^f/CLIP-I score. The blue number indicates the proposed method outperforms its baseline counterpart, while the red number indicates the opposite. The GAP is the average improvement over all concepts. Qualitative results are shown in Fig. 17.

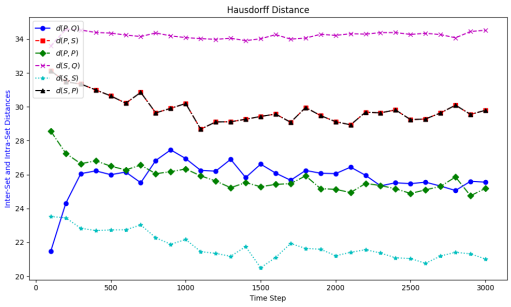
Method	dog	cat	tort*	teddy*	chair	table	flower	wpot*	barn	GAP
TI	26.05/ 60.18	26.64/ 79.65	30.49/ 73.09	27.29/ 77.05	27.19/ 77.97	27.06/ 64.36	25.26/ 68.63	28.0/ 67.02	27.45/ 80.22	0.00/ 0.00
TI+TEA	26.14/ 58.97	27.94/ 77.47	30.71/ 72.74	27.68/ 76.02	28.13/ 73.27	27.78/ 61.33	26.28/ 63.77	28.34/ 66.44	28.2/ 77.06	0.64/ -2.34
DB	20.28/ 59.93	20.29/ 74.82	24.55/ 91.14	18.89/ 85.52	21.87/ 87.5	20.77/ 73.83	20.07/ 82.98	20.43/ 66.69	23.58/ 80.98	0.00/ 0.00
DB+TEA	21.71/ 66.25	20.1/ 92.43	25.08/ 91.1	19.65/ 86.98	23.61/ 86.6	22.97/ 84.87	23.12/ 76.05	24.99/ 82.46	26.31/ 77.77	1.87/ 4.57
CD	27.33/ 56.09	29.37/ 78.34	31.33/ 78.42	28.47/ 77.67	27.16/ 69.88	27.88/ 62.84	26.7/ 62.66	30.0/ 69.18	26.34/ 70.95	0.00/ 0.00
CD+TEA	27.45/ 56.18	29.25/ 77.59	31.06/ 79.61	28.04/ 78.64	28.16/ 71.81	28.34/ 63.08	27.37/ 62.63	30.15/ 68.0	27.82/ 71.78	0.34/ 0.37

Table 7: Results on CelebA dataset. The first/second metric is the CLIP_T^f/CLIP-I score. The blue number indicates the proposed method outperforms its baseline counterpart, while the red number indicates the opposite. The GAP is the average improvement over all concepts. Qualitative results are shown in Fig. 16.

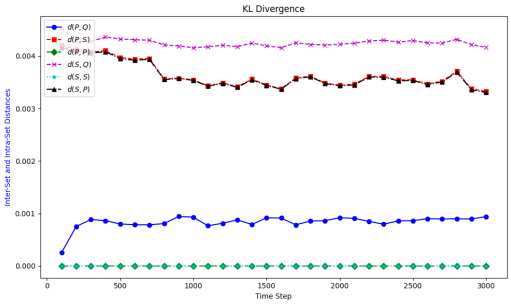
Method	124	181	276	342	351	437	615	908	1335	1429	GAP
TI	23.7/ 68.93	26.29/ 61.68	21.32/ 65.33	26.06/ 75.64	25.25/ 69.63	23.46/ 71.19	26.54/ 74.37	24.61/ 64.5	25.4/ 67.55	24.97/ 60.88	0.00/ 0.00
TI+TEA	24.23/ 68.48	26.49/ 59.6	20.97/ 65.2	27.28/ 71.01	26.11/ 65.73	23.93/ 66.25	26.89/ 70.56	25.98/ 62.97	25.69/ 66.72	25.68/ 59.01	0.57/ -2.41
DB	25.37/ 54.78	25.91/ 61.75	22.11/ 58.7	24.37/ 70.41	25.0/ 62.95	20.65/ 75.15	25.06/ 66.56	24.93/ 63.46	23.66/ 59.2	24.65/ 60.42	0.00/ 0.00
DB+TEA	26.26/ 50.67	26.59/ 60.23	24.3/ 47.66	25.58/ 65.83	25.43/ 61.34	22.56/ 68.79	26.03/ 60.88	25.92/ 60.34	24.79/ 53.66	25.37/ 58.46	1.11/ -4.55
CS	26.92/ 43.56	26.87/ 52.48	26.91/ 36.22	27.14/ 52.64	26.94/ 53.39	26.97/ 55.45	26.99/ 53.96	26.93/ 53.32	27.05/ 44.95	26.88/ 50.33	0.00/ 0.00
CS+TEA	27.09/ 44.88	26.83/ 55.03	27.18/ 37.1	27.37/ 53.63	26.99/ 55.56	27.11/ 56.41	27.16/ 55.01	26.93/ 55.33	26.94/ 46.68	26.98/ 52.27	0.09/ 1.56



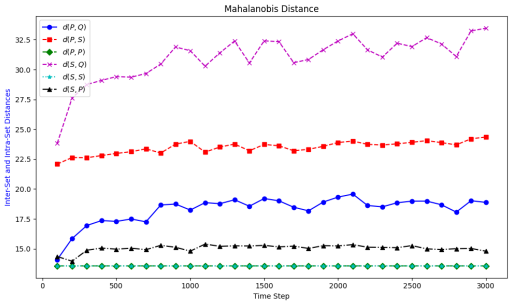
(a) L2 distance



(b) Hausdorff distance



(c) KL divergence



(d) Mahalanobis distance

Figure 10: Different distance metrics between the sets $P = \{\tau(\lfloor a_i, V^* \rfloor)\}$ and $Q = \{\tau(\lfloor a_i, c \rfloor)\}$ in Textual Inversion (TI), showing the semantic shifting of the learned embedding over training iterations.

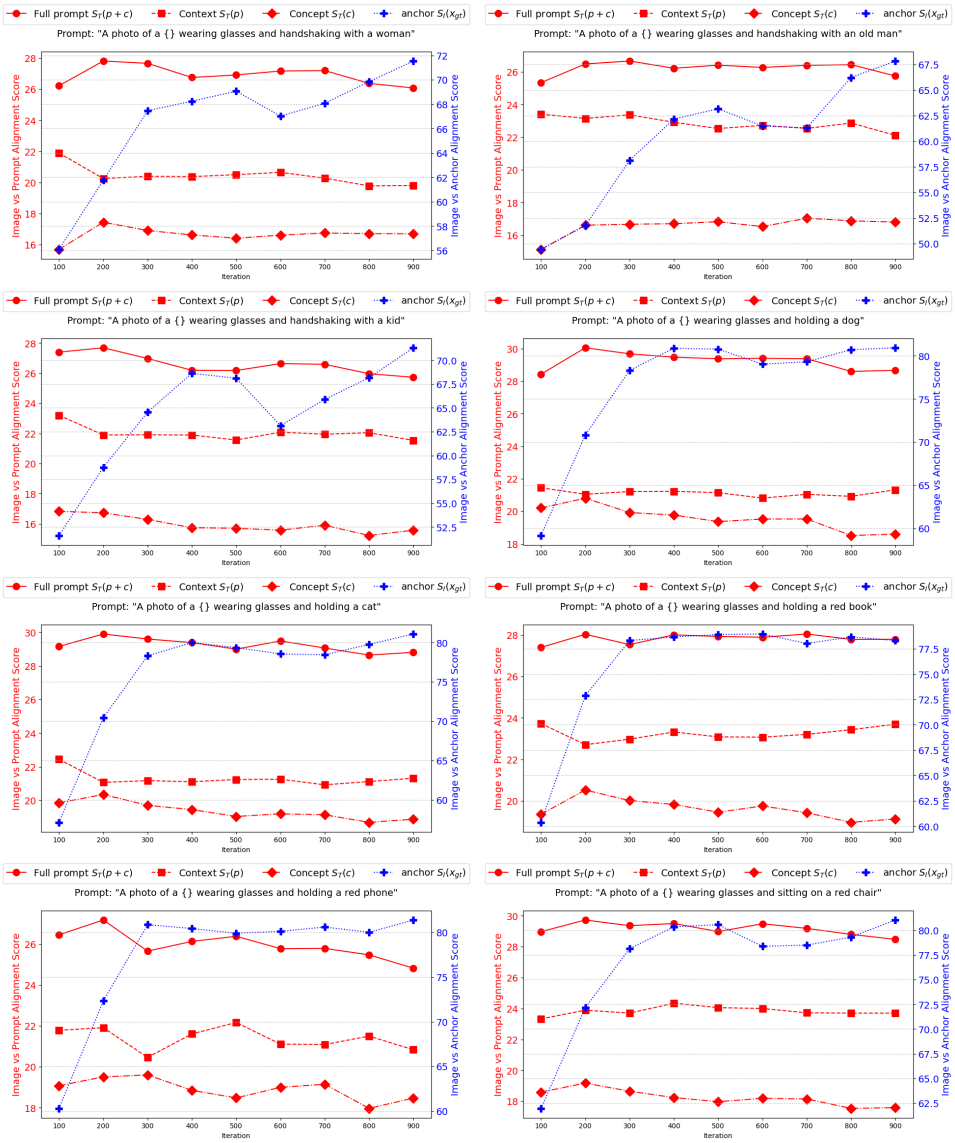


Figure 11: Alignment scores showing the SCP on Textual Inversion with different prompts.

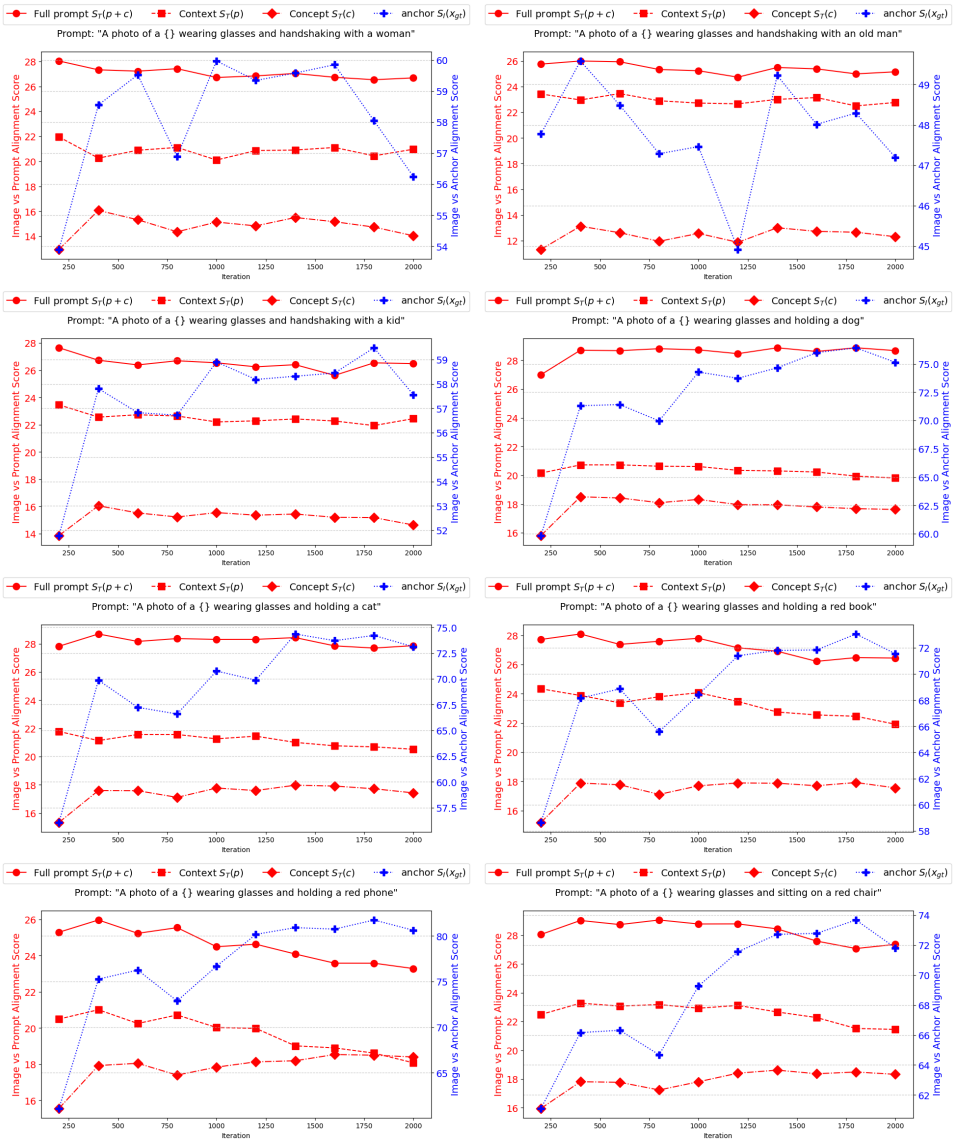
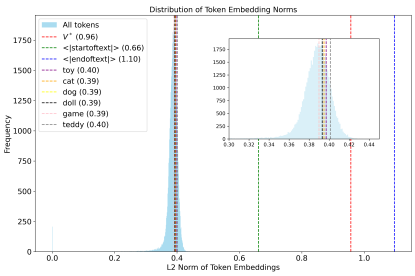
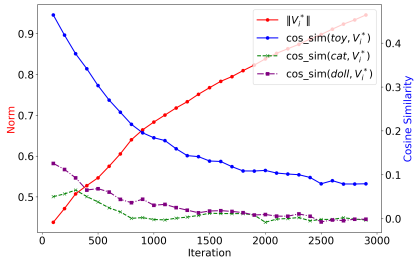


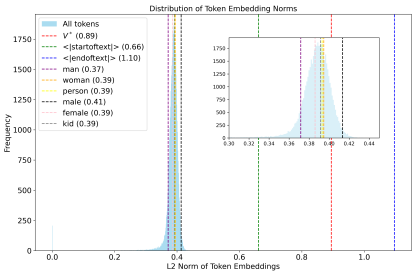
Figure 12: Alignment scores showing the SCP on DreamBooth with different prompts.



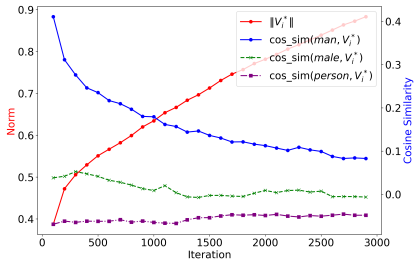
(a) Distribution of the norm in the cat-toy setting



(b) The semantic shift in the cat-toy setting



(c) Distribution of the norm in the CelebA setting



(d) The semantic shift in the CelebA setting

Figure 13: Left: The distribution of the norm of the token embedding M including special token V^*, Right: The semantic drift of V^* in term of magnitude and direction over time. The same phenomenon is observed in DreamBooth as shown in Figure 14.

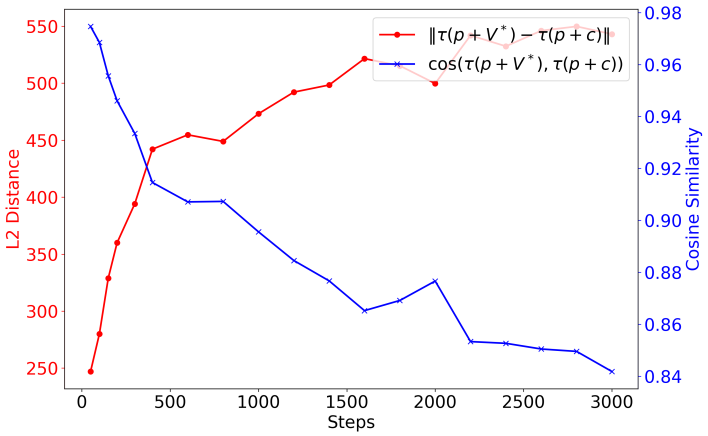


Figure 14: The semantic drift of the embedding of entire prompt $\lfloor p, V^* \rfloor$ in term of magnitude and direction over time with DreamBooth.

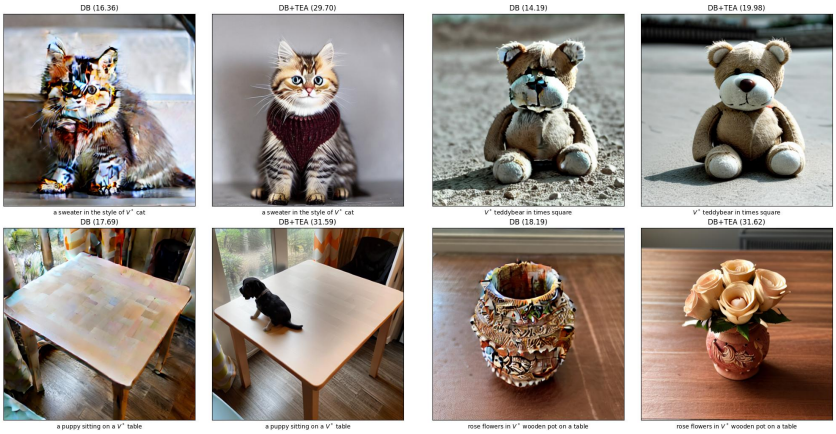


Figure 15: Some cool effects of our method (DB+TEA) over the baseline counterpart DB. The SCP in DB lead to the distorted generated images. Our method successfully corrects the semantic embedding and generates more coherent and realistic images.

(a) 124

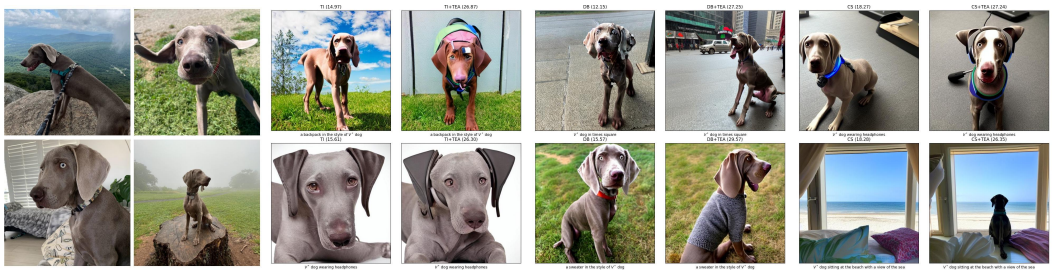
(b) 181

(c) 342

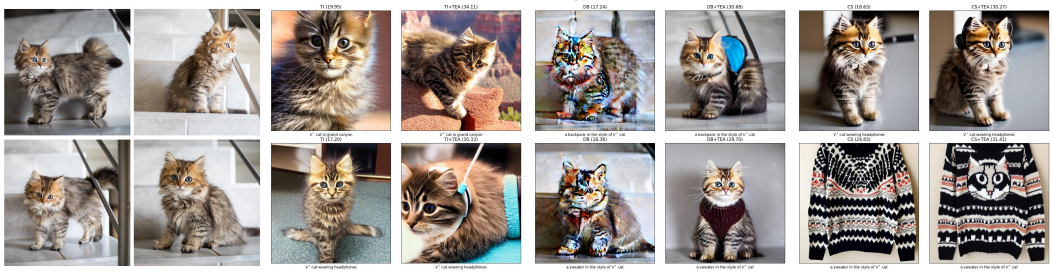
(d) 437

Figure 16: Qualitative comparison between baseline methods and their TEA variants on the CelebA dataset. Column 1-2: Reference images. Column 3: TI, Column 4: TI+TEA, Column 5: DB, Column 6: DB+TEA. Input prompts are shown below each image while alignment scores are shown on the top. More results can be found in the anonymous repository.

(a) DecorItems Wooden Pot



(b) Pet Dog



(c) Pet Cat

(d) Scene Barn

Figure 17: Qualitative comparisons between baseline methods and their TEA variants on the Custom-Concept101 dataset. Column 1-2: Reference images. Column 3: TI, Column 4: TI+TEA, Column 5: DB, Column 6: DB+TEA. Input prompts are shown below each image while alignment scores are shown on the top. More results can be found in the anonymous repository.

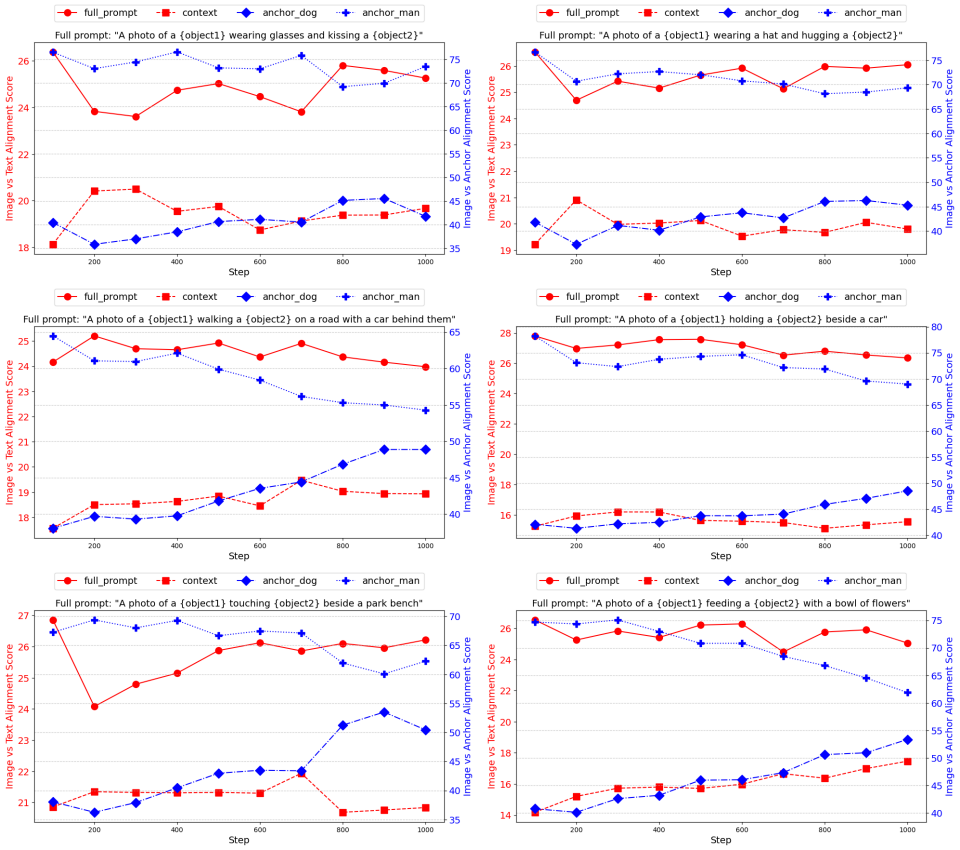
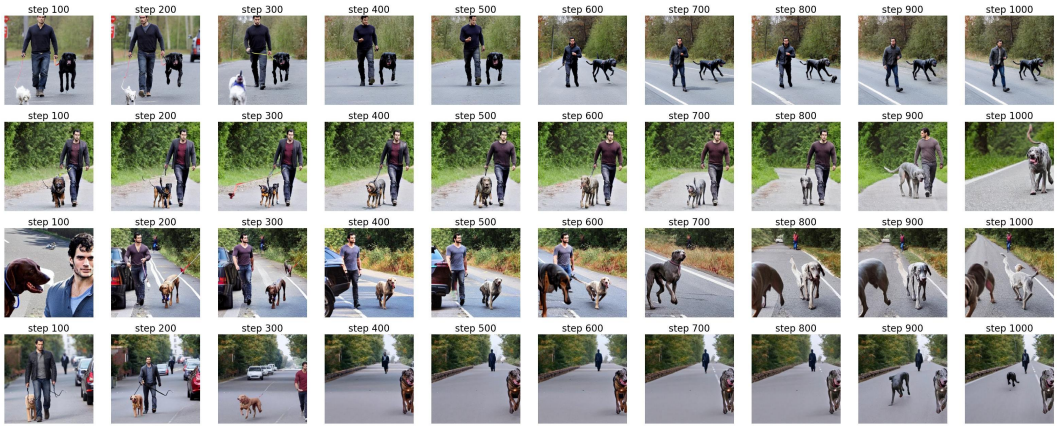
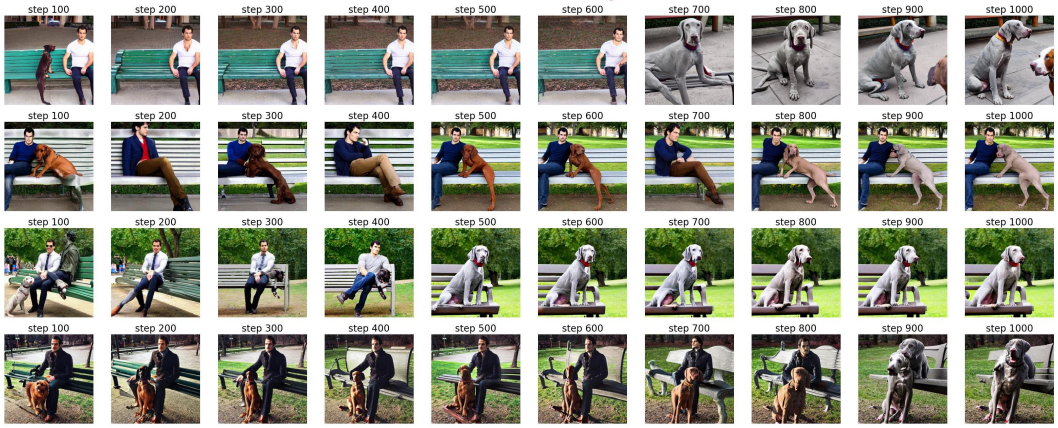


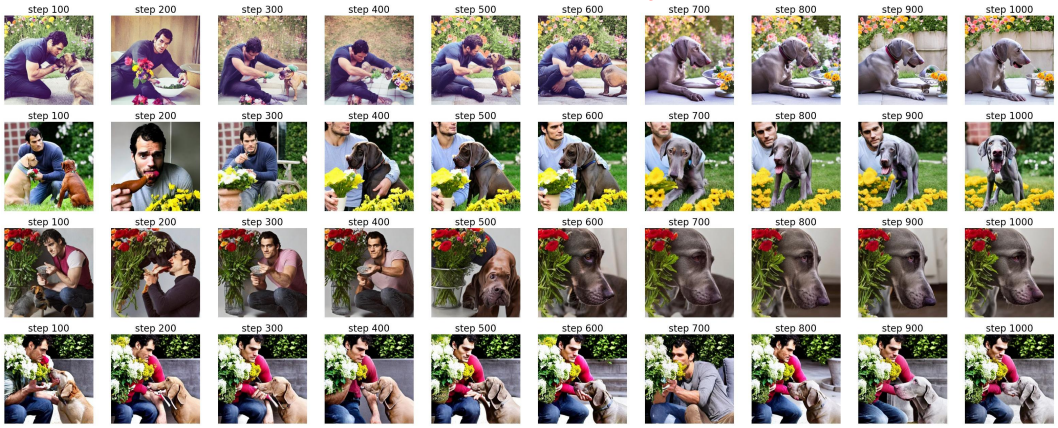
Figure 18: Alignment scores showing the SCP in multiple-concept personalization settings. The embedding of V^*_{man} held fixed while the embedding of V^*dog is varied over fine-tuning step. See Figures 19 and 20 for the example images.



(a) Prompt: 'A photo of a V^*_{man} walking a V^*_{dog} on a road with a car behind'



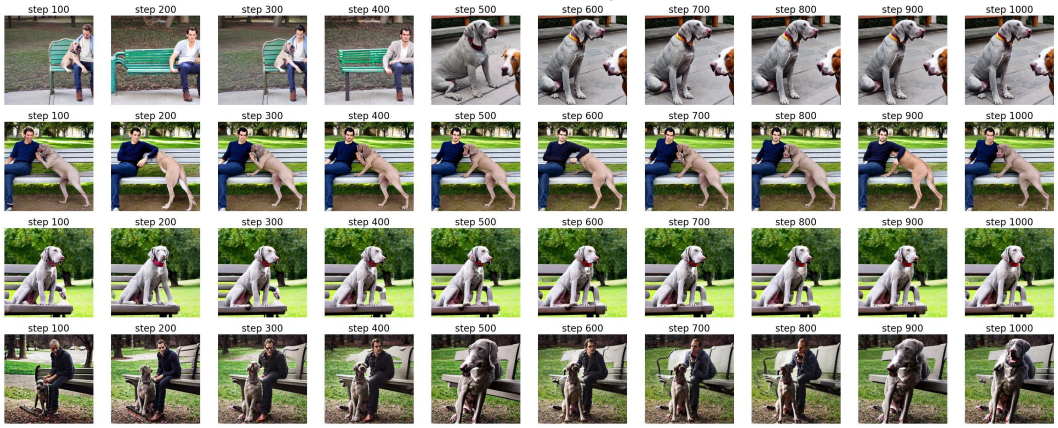
(b) Prompt: 'A photo of a V^*_{man} touching a V^*_{dog} beside a park bench'



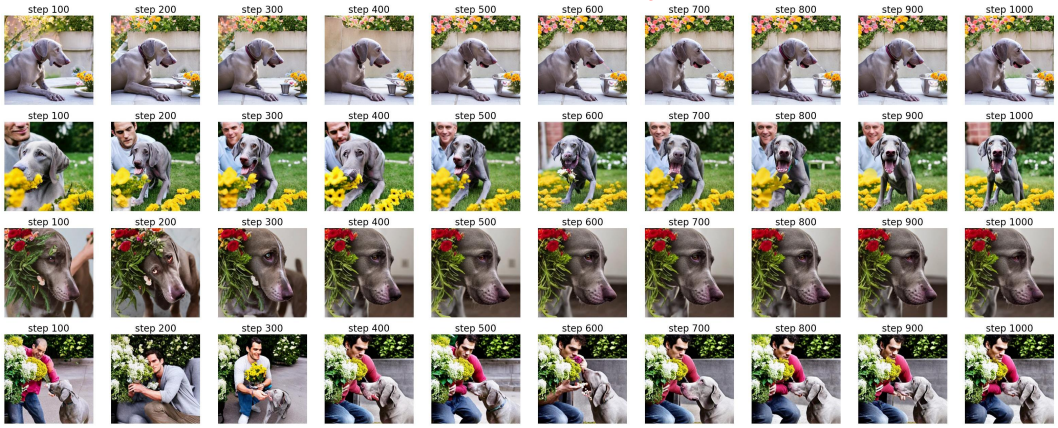
(c) Prompt: 'A photo of a V^*_{man} feeding a V^*_{dog} with a bowl of flowers'

Figure 19: Illustration of the semantic collapsing problem on multi-concepts personalization. The embedding of V^*_{man} **is fixed** while the embedding of V^*_{dog} **is varied** across different training steps. See Figure 5 for the detailed alignment scores.

(a) Prompt: 'A photo of a V^*_{man} walking a V^*_{dog} on a road with a car behind'



(b) Prompt: 'A photo of a V^*_{man} touching a V^*_{dog} beside a park bench'



(c) Prompt: 'A photo of a V^*_{man} feeding a V^*_{dog} with a bowl of flowers'

Figure 20: Illustration of the semantic collapsing problem on multi-concepts personalization. The embedding of V^*_{man} **is varied** across different training steps while the embedding of V^*_{dog} **is fixed**. See Figure 5 for the detailed alignment scores.