TOPA: Mitigating Concept Dominance in Diffusion Personalization via Target-Oriented Perturbation Augmentation

Jiayou Lu[1*], Mingzhi Lyu[2*], Junfeng Huang[2], and Adams Wai-Kin Kong[2*]

[1] Energy Research Institute @ NTU (ERI@N), Nanyang Technological University, Singapore

[2] College of Computing and Data Science, Nanyang Technological University, Singapore

Abstract. Concept personalization injects a user-specified concept into a pretrained diffusion model from only a few reference images, enabling customized content creation. However, existing personalization methods frequently suffer from concept-dominant failures—while the personalized concept is well preserved, other prompt-specified contexts (e.g., background, attributes, and interactions) are not properly generated, thereby limiting controllability. To address this problem, in this paper, we propose **T**arget-**O**riented **P**erturbation-based **A**ugmentation (TOPA), a reference images-only augmentation framework. TOPA optimizes additive perturbations on reference images to provide token–region attention guidance and further reduces background entanglement via subject-isolation compositing. Unlike prior solutions that modify training objectives or architectures, TOPA requires no changes to personalization pipelines, and its augmented datasets can be directly applied to existing methods. Extensive experiments across multiple personalization methods demonstrate consistent improvement on context adherence while preserving concept fidelity.

Keywords: Data augmentation · Concept Personalization · Diffusion model

1 Introduction

Diffusion-based generative models have become a dominant paradigm for few-shot concept personalization, enabling a new visual subject to be learned from only a handful of images. [1] A broad family of methods—including Textual Inversion (TI) [5], DreamBooth [19], Custom Diffusion [10], ELITE [25], and other

* Equal contribution, may cite either first

[1] In this paper, we use "subject" to denote the localized visual entity in reference images to be personalized (e.g., the view of a specific toy), and "concept" to denote its abstract knowledge that can be learned as representation in the diffusion models (e.g., a special token embedding or other personalization parameters).

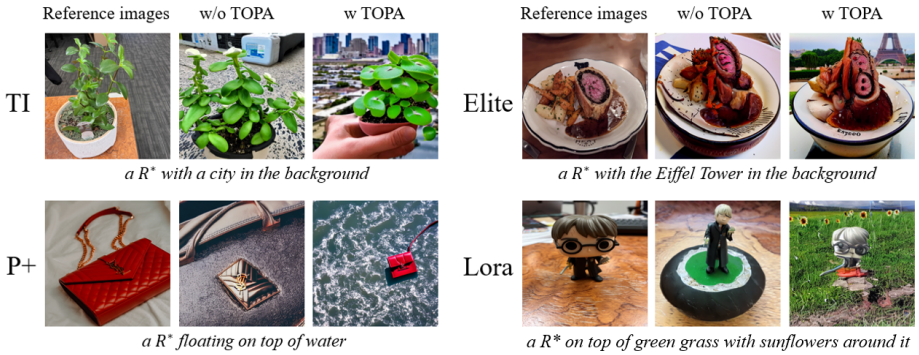

Fig. 1: Examples of concept-dominance failures and the effect of TOPA. For each example, the left column shows the reference images. The middle column shows generations from the corresponding personalization method *without* TOPA, where the learned concept dominates and suppresses prompt-specified context. The right column shows generations *with* TOPA, which restores context controllability and improves compositionality. R^* in the prompts represents the learned concept token.

variants—inject concept-specific knowledge into a pretrained diffusion model with minimal data and computation, thereby supporting personalized content creation and controllable generation. Concretely, once a user personalizes a concept (e.g., a particular toy) and associates it with a special token such as `<toy_one>`, the model can generate diverse images containing the same concept under different prompts, rather than producing only generic instances.

Despite this progress, existing personalization methods commonly suffer from a critical limitation that we term the **concept-dominant failure**. After personalization, generations conditioned on the learned concept token often preserve the subject faithfully, yet under-realize prompt-specified context (e.g., scene, attributes, interactions), undermining compositional generation. TI provides a representative example: It learns a concept by optimizing a single token embedding, and can yield high subject fidelity. However, as illustrated in Fig. 1, the personalized token may overwhelm other prompt semantics, producing images that focus on the subject while failing to render the desired context. This issue has been recognized in recent studies [1,30], which propose training-time modifications or method-specific regularization to improve compositionality. Nevertheless, many such solutions [26,30] are tightly coupled to particular personalization algorithms (often TI), or require users to alter their local training pipelines, limiting scalability and large-scale deployment. In contrast, data-level solutions that require minimal pipeline changes and generalize across diverse personalization methods remain under-explored.

To address these failures, we propose Target-Oriented Perturbation-based Augmentation (TOPA), an offline, reference images-only augmentation framework designed to mitigate concept dominance without modifying model architectures or personalization objectives. **To the best of our knowledge, TOPA is the first augmentation method that explicitly targets concept domi-**

nance in diffusion personalization via optimization-based token–region alignment. Our key insight is that personalization often suffers from cross-token interference in cross-attention (CA). Context tokens undesirably allocate attention to the subject region, which suppresses their ability to control background attributes and leads to the concept-dominance failure. TOPA corrects this by optimizing imperceptible, region-controlled perturbations on the reference images to strengthen token–region alignment under cross-attention. Moreover, many personalization methods, including TI and Custom Diffusion, rely on the standard diffusion training objective and lack explicit supervision indicating which pixels correspond to the target subject. In the low-shot regime, the model can therefore inadvertently encode spurious correlations between the subject and co-occurring backgrounds, especially when the reference dataset exhibits limited background diversity. To reduce such subject–background entanglement under TOPA, we introduce a concept-isolation pipeline that increases background diversity via realistic compositing, encouraging the personalized token to capture the subject rather than background cues.

Our contributions are summarized as follows:

- We propose TOPA, an offline augmentation framework that strengthens concept binding by optimizing imperceptible, region-controlled perturbations to improve token–region cross-attention alignment, without changing personalization algorithm or objectives.
- We introduce a subject-isolation compositing pipeline to increase background diversity and reduce spurious correlations, improving compositional generalization after personalization.
- Extensive experiments across multiple personalization methods demonstrate that TOPA consistently mitigates concept dominance and is complementary to existing regularization-based approaches.

2 Related Work

2.1 Diffusion-based Concept Personalization

Diffusion models have enabled effective few-shot concept personalization by adapting pretrained text-to-image generators to novel concepts using only a handful of subject-specific reference images. A representative line of work focuses on *token-level* customization, where the diffusion backbone remains frozen while the textual conditioning is specialized. Textual Inversion (TI) [5] learns a dedicated embedding for a newly introduced pseudo-token, allowing the token to represent a user-provided subject with minimal training data. Beyond single-token embeddings, several works improve expressiveness by expanding the conditioning space. P+ [8] introduces an extended textual conditioning scheme that learns different token embeddings for different cross-attention layers in the U-Net, improving both capacity and controllability of the personalized representation.

Another line of work personalizes diffusion models by fine-tuning a subset of model parameters. DreamBooth [19] fine-tunes a pretrained diffusion model to associate a special token with the target concept, and employs a prior-preservation objective to reduce overfitting while retaining the model's language prior. Custom Diffusion [10] further shows that updating only a small set of parameters—typically in cross-attention-related components—can efficiently bind a new concept while keeping most of the model fixed.

In addition, encoder-based methods aim to avoid optimization for each concept at test time. ELITE [25] learns an image-to-text embedding encoder that produces customized textual embeddings from reference images, and injects patch-level information into cross-attention to improve fidelity and editability.

2.2 Compositionality and Regularization in Personalization

A growing body of work studies the trade-off between subject fidelity and compositionality in personalized generation. Existing approaches typically modify training objectives, introduce regularization, or adjust update rules to improve the binding between concept tokens and visual content. For instance, DreamBooth's prior-preservation objective leverages class-level priors as a self-supervised regularizer to maintain generative diversity when fine-tuning on small subject-specific datasets. Compositional Inversion [30] regularizes newly introduced embeddings using pretrained token embeddings as anchors, encouraging them to remain close to the original embedding distribution. Similarly, CoRe [26] improves textual embeddings by constraining the embeddings and cross-attention maps of context tokens to remain consistent with a reference prompt where the special token is replaced by an existing one, helping prevent semantic distortion and improving text alignment.

Some methods instead modify the generation procedure for finer control. To address the issue that certain subject tokens may be ignored during denoising, Chefer et al. [1] propose Attend-and-Excite to optimize the latent at each step to increase the cross-attention score of the most neglected token. A more recent work further extends classifier-free guidance with adaptive variants to better balance compositionality and subject fidelity during sampling [17].

While effective, many of these strategies are *method-specific* and require modifying training or inference pipelines (e.g., loss functions, update scope, or additional constraints), which may limit scalability and compatibility across personalization algorithms. In contrast, our work addresses compositional failures caused by concept dominance through a reference-image-only augmentation strategy, requiring no modification to model architectures, personalization objectives, or sampling procedures.

2.3 Data Augmentation for Diffusion Models

Data-centric strategies are widely used to improve generalization and robustness in vision systems. Beyond basic transformations (e.g., cropping, flipping, and color jittering), learned augmentation policies automatically discover effective

transformations from data [3, 4, 16]. Sample-mixing strategies such as MixUp and CutMix further enrich training distributions by combining multiple samples and labels [7, 28, 29].

In personalization methods, augmentation is typically used to increase training diversity, such as random rescaling in Custom Diffusion [10]. A few studies explore background diversification for subject-centric learning. For instance, BLIP-Diffusion [11] uses background-replaced images as inputs to a multi-modal encoder to extract embeddings that are less entangled with background; the original training images are then used to learn concept representations conditioned on these embeddings. Although such results highlight the importance of background diversity, *explicitly* augmenting personalization data with mechanisms guided by cross-attention to disentangle subject and context remains underexplored.

Separately, perturbation-based image modifications have been extensively studied in adversarial robustness and data poisoning for diffusion models, including TrojDiff [2], MIST [13], SZT [21], and Diff-Protect [27]. These works demonstrate that carefully designed input perturbations can substantially influence, or even corrupt, learned representations. Inspired by the effectiveness of optimization-based perturbations—but targeting a different goal—we propose to optimize imperceptible perturbations on personalization training images to mitigate concept dominance and improve concept–context compositionality.

3 Method

3.1 Preliminaries

Current personalization methods typically build on Latent Diffusion Models (LDMs) due to their efficiency and modularity compared to pixel-space diffusion models. In LDMs, an input image x is first encoded into a latent representation $z_0 = \mathcal{E}(x)$ using a pretrained variational autoencoder (VAE). The latent is then perturbed through a forward diffusion process:

$$z_t = \sqrt{\overline{\alpha}_t}\, z_0 + \sqrt{1-\overline{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\overline{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. A U-Net is trained to predict the added noise ϵ given the noisy latent z_t, timestep t, and a conditioning vector $\mathbf{c}_{\theta'}$, typically derived from a text encoder:

$$\arg\min_{\theta,\theta'} \mathbb{E}_{x,t,\epsilon}\left[\left\|\epsilon_\theta(z_t, t, \mathbf{c}_{\theta'}) - \epsilon\right\|_2^2\right]. \tag{2}$$

After training, samples are generated by reversing this process from random noise $z_T \sim \mathcal{N}(0, I)$, followed by decoding with the VAE decoder $\mathcal{D}(z_0)$.

Different personalization methods optimize different subsets of the model. For instance, TI learns only a new embedding θ_{R^*} for a pseudo-token R^*, keeping all other parameters frozen. Given prompts y^* containing R^*, TI minimizes:

$$\arg\min_{\theta_{R^*}} \mathbb{E}_{x_f,t,\epsilon}\left[\left\|\epsilon_\theta(z_t, t, \tau_{\theta'}(y^*)) - \epsilon\right\|_2^2\right], \tag{3}$$

where $\tau_{\theta'}(y^*)$ is the encoded vector of the prompt. This embedding allows the model to associate the new concept with arbitrary contexts during generation.

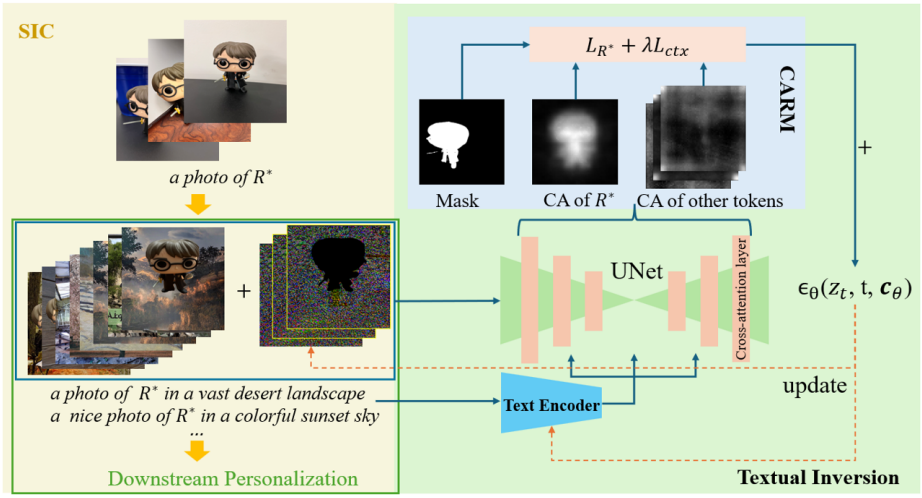

Fig. 2: Overview of TOPA. Given a few-shot reference set, Subject-Isolation Compositing (SIC) diversifies backgrounds and the corresponding background prompts by compositing the segmented subject onto synthesized backgrounds. Cross-Attention Rebalancing Module (CARM) then learns imperceptible, region-specific perturbations to rebalance token-level cross-attention during Textual Inversion. The resulting augmented reference images are finally used for downstream diffusion personalization.

3.2 Cross-Attention Rebalancing Module

Most personalization methods optimize on a small set of reference images without explicitly disentangling the target subject from its co-occurring background. As observed in prior work, the denoising objective may concentrate learning signal on the common subject regions shared across reference images [21], while leaving the allocation of cross-attention between the newly introduced concept token and other context tokens largely unconstrained. In practice, this often induces *cross-token interference*: context tokens waste attention on subject pixels, reducing their ability to control background attributes, and aggravating the concept-dominance issue during generation. For clarity, we describe the formulation using Textual Inversion (TI) as a representative example.

Overview. We propose to learn a small additive perturbation δ for each reference image x to *rebalance* cross-attention during TI. Intuitively, the perturbation is optimized such that (i) the pseudo-token embedding e^* primarily binds to the subject region, and (ii) context tokens are encouraged to attend to background regions, thereby reducing context-to-subject interference. Let the prompt be y^* containing the pseudo-token R^* with embedding e^*, and denote the token embeddings of y^* as $\hat{e} = \{e_0, e_1, \ldots, e^*, \ldots, e_N\}$, which are encoded by a text encoder

to obtain the conditioning vector $\tau_{\theta'}(y^*)$. Given a segmented subject mask M for the concept region, we optimize δ on the augmented image $\tilde{x} = x + \delta$ as shown in Fig. 2.

Aggregated cross-attention maps. Let $A_c(\tilde{x}, y^*)$ denote the aggregated cross-attention map for token c produced by the U-Net, where we average attention over a set of selected layers and heads. We define the context token set $\mathcal{C}$ as all tokens in y^* except the pseudo-token R^* and stopwords, and compute the average context attention map:

$$A_{\text{ctx}}(\tilde{x}, y^*) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} A_c(\tilde{x}, y^*). \tag{4}$$

Context-to-subject interference suppression. Our first objective discourages context tokens from attending to the subject region and encourages them to focus on the background. Formally, we penalize the overlap between the subject mask M and the context attention map:

$$\mathcal{L}_{\text{ctx}} = \| M \odot A_{\text{ctx}}(x + \delta, y^*) \|_2^2 , \tag{5}$$

This term directly targets the primary failure mode behind concept dominance: when context tokens allocate attention to the subject, they lose controllability over background semantics, resulting in suppressed or missing context attributes in generated images.

Subject-token anchoring. Optimizing $\mathcal{L}_{\text{ctx}}$ alone, however, is under-constrained in a competitive nature of attention. Although pushing context tokens toward the background can reduce cross-token interference, it does not explicitly anchor the pseudo-token R^* to the subject region; consequently, the pseudo-token attention may spread spatially as a compensatory behavior. To stabilize the subject binding and encourage a consistent partition of labor between subject and context, we additionally align the pseudo-token attention with the subject mask:

$$\mathcal{L}_{R^*} = \text{MSE}\Big(M,\ A_{R^*}(x + \delta, y^*)\Big), \tag{6}$$

where $A_{R^*}(\cdot)$ denotes the aggregated cross-attention map of the pseudo-token.

Empirically, we find that the joint objective provides a more reliable concept–context balance than optimizing either term alone, as it simultaneously (i) suppresses context-to-subject interference and (ii) anchors the pseudo-token to the subject region. We summarize the proposed cross-attention rebalancing objective as:

$$\mathcal{L}_{\text{CARM}} = \mathcal{L}_{\text{ctx}} + \lambda \, \mathcal{L}_{R^*}, \tag{7}$$

where λ controls the trade-off between interference suppression and subject-token anchoring.

Alternating perturbation–embedding optimization. A key challenge in optimizing the perturbations $\{\delta_i\}$ for reference image $\{x_i\}$ is that the pseudo-token embedding e^* is *not* static during personalization. Since the augmented images are repeatedly used throughout TI training, the attention maps and

Algorithm 1 Alternating Perturbation–embedding Optimization

Require: Reference images $\{x_i\}_{i=1}^{M}$, masks $\{M_i\}$, prompt template set $\mathcal{P}$, update interval K, total steps S

Initialize $\delta_i \leftarrow 0$ for all i
Initialize pseudo-token embedding e^* as in standard TI
for $s = 1$ to S **do**
    Sample index $i \sim \{1, \ldots, M\}$
    Sample prompt $y^* \sim \mathcal{P}$
    $\tilde{x}_i \leftarrow \text{clip}(x_i + \delta_i)$
    $z_0 \leftarrow \mathcal{E}(\tilde{x}_i)$
    Sample diffusion timestep τ and noise ϵ
    $z_\tau \leftarrow \sqrt{\bar{\alpha}_\tau} z_0 + \sqrt{1 - \bar{\alpha}_\tau}\epsilon$
    $\epsilon_{\text{pred}} \leftarrow \epsilon_\theta(z_\tau, \tau, \tau_\theta(y^*))$
    if $s \bmod K = 0$ **then**
        $\mathcal{L} \leftarrow \mathcal{L}_{\text{CARM}}$
        Update δ_i using $\nabla\mathcal{L}$
    else
        $\mathcal{L} \leftarrow \|\epsilon_{\text{pred}} - \epsilon\|_2^2$
        Update e^* using $\nabla\mathcal{L}$
    end if
end for
return $\{\tilde{x}_i\}$ for final TI training

$\mathcal{L}_{\text{CARM}}$ depend on the evolving embedding e^*. If we optimize δ with a fixed embedding (e.g., the initialization), the resulting perturbations may overfit to an early-stage token representation and become suboptimal for later training stages. To address this coupling, we adopt an alternating update strategy that jointly refines δ and e^* during a dedicated perturbation learning stage, as illustrated in Algorithm 1.

Specifically, we perform standard TI updates on e^*, while updating the perturbation δ_i less frequently (every K steps). This schedule stabilizes optimization by allowing e^* to adapt to the current augmented samples before δ is updated again. After obtaining the perturbations, we can discard the intermediate embedding from this stage and perform a standard personalization training (TI or other personalization methods) on the augmented dataset $\{\tilde{x}_i\}$ to learn the final concept used for downstream tasks.

3.3 Subject-Isolation Compositing

While CARM rebalances cross-attention during optimization, it does not alter the intrinsic statistical correlation between the subject and background in the reference dataset. In typical personalization scenarios, the limited reference images often share similar background patterns, which may bias the pseudo-token embedding toward background-correlated features. To further enhance subject–background disentanglement, we introduce a data-level augmentation strategy termed **Subject-Isolation Compositing (SIC)**.

Background diversification. The key idea of SIC is to increase background diversity while preserving subject identity. Given a concept with h reference images, we first construct a set of q diverse background prompts that describe different background details. Using a pretrained Stable Diffusion model, we synthesize q background images corresponding to these prompts. For each reference image, we segment the subject region using the same automatic segmentation pipeline employed in Sec. 3.2, and composite the extracted subject onto each generated background image. As a result, we obtain $h \times q$ composited images per concept. This compositing strategy amplifies background variability while keeping the subject appearance fixed. Consequently, during personalization, the subject becomes the only consistent feature across the augmented dataset, encouraging the trainable pseudo-token embedding to focus on concept-specific attributes rather than spurious background correlations.

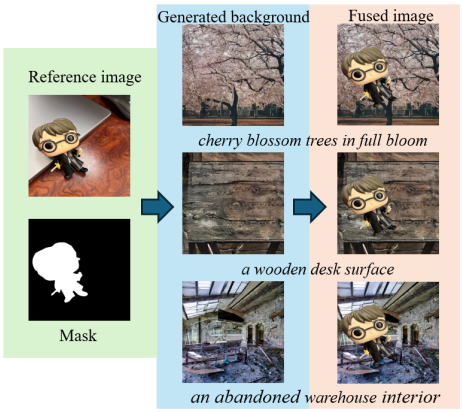

Fig. 3: Overview of SIC. The subject is segmented from the reference image and composited onto backgrounds generated from diverse prompts.

Region-based perturbation optimization. Although SIC substantially enriches background diversity, directly applying Algorithm 1 to all $h \times q$ composited images would be computationally expensive, since CARM requires optimizing a perturbation for each image individually.

To improve efficiency, we exploit the structural decomposition of composited images into subject and background regions. Specifically, each composited image can be expressed as:

$$\tilde{x}_{i,j} = \mathcal{C}(s_i, b_j), \tag{8}$$

where s_i denotes the segmented subject from the i-th reference image, b_j denotes the j-th generated background, and $\mathcal{C}(\cdot,\cdot)$ denotes the compositing operator. An example is illustrated in Fig. 3.

Multiple composited images share identical subject regions s_i or identical background regions b_j. Therefore, instead of optimizing an independent perturbation for each $\tilde{x}_{i,j}$, we adopt a *region-based perturbation strategy*:

- **Subject-region perturbation:** Learn a universal perturbation $\delta_i^{(s)}$ embedded only within subject region s_i, shared across all backgrounds b_j.
- **Background-region perturbation:** Learn a universal perturbation $\delta_j^{(b)}$ embedded only within background region b_j, shared across all subjects s_i.

With either above sharing scheme, the number of perturbation optimizations reduces from $h \times q$ to either h (subject-level sharing) or q (background-level shar-

ing), depending on which region the perturbation is applied to. This significantly lowers computational cost while maintaining the disentanglement benefits of SIC.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate our method on 10 concepts randomly selected from CustomConcept101 [10], following similar dataset scale as related research [1, 6, 8, 15, 30]. We also present quantitative evaluations on more concepts and qualitative evaluations of image generation in Appendix. Each concept in CustomConcept101 contains 3–15 reference images depicting the same subject under varying viewpoints and backgrounds. Following the standard personalization protocol, these few-shot reference images are used to learn a dedicated pseudo-token or a small subset of model parameters for each concept, depending on the specific personalization method.

Personalization methods. We test TOPA on five representative personalization approaches, including Textual Inversion (TI) [5], DreamBooth LoRA [19], Custom Diffusion [10], ELITE [25], and P+ [8]. Unless otherwise specified, all methods are implemented using their official or widely adopted configurations, with the same diffusion backbone for fair comparison.

Evaluation protocol. For each concept, we first apply TOPA to its reference image set. We then train a personalized concept using the corresponding personalization method. Finally, we evaluate compositional generalization by generating images using 24 prompts from DreamBench [19] under diverse contextual descriptions. For each prompt, we generate 10 images and report results aggregated over all prompts and concepts, resulting in 240 images per concept. All experiments use fixed random seeds for augmentation, personalization training, and inference to ensure reproducibility and fair comparison.

Diffusion backbone. Personalization methods and TOPA are built and trained upon Stable Diffusion v1.4 as the pretrained latent diffusion backbone. We also present the results of applying TOPA trained with Stable Diffusion v1.4 on Stable Diffusion v3, v3.5, FLUX.1, and FLUX.2 in Appendix.

TOPA augmentation details. For each concept, TOPA consists of two stages. (1) Subject-Isolation Compositing (SIC): We diversify backgrounds by replacing the background of each reference image with a set of diffusion-generated backgrounds [2], producing 34 composited variants per reference image. The prompts to generate backgrounds are presented in Appendix. (2) Cross-Attention Rebalancing Module (CARM): We further optimize region-specific perturbations on the composited images, where perturbations are applied to the *background region* by default. λ to balance $\mathcal{L}_{\text{ctx}}$ and $\mathcal{L}_{R^*}$ is set as 1.0. During the perturbation learning stage, we set $S = 5,000$ and $K = 2$, that is, we optimize the pseudo-token embedding for 2,500 iterations with a learning rate 1×10^{-4} for embedding

[2] There is no overlap between the prompts used to generate the backgrounds for SIC and those used to generate images in personalization evaluation.

updating, and update perturbations for 2,500 iterations with a learning rate of 0.02 following the alternating scheme in Sec. 3.

4.2 Evaluation Metrics

We evaluate personalization performance from four aspects: image quality, distribution alignment, background controllability, and concept fidelity.

Image quality. We measure perceptual quality using BRISQUE [14] and CLIP-based Image Quality Assessment (CLIP-IQA) [23]. BRISQUE is a no-reference metric that reflects natural image statistics, while CLIP-IQA estimates perceptual quality through CLIP-based scoring.

Distribution alignment. We report Fréchet Inception Distance (FID) [9] between generated images and the corresponding reference distribution to assess overall distribution consistency. Lower FID indicates better alignment with the target concept distribution.

Background controllability via VQA and text–image alignment. To quantify prompt-driven background adherence, we leverage three independent vision-language models (Flan [24], BLIP [12], and LLaMA [22]) to perform visual question answering (VQA) on generated images. For each input prompt, we construct questions to ask about background presence and compute the agreement between predicted answers and the prompt-specified background semantics. We also report CLIP text similarity between prompts and text descriptions from BLIP describing the background of generated images, denoted as BLIP-VQA-Sim. The detailed explanation of these settings can be found in Appendix. In addition, we report CLIP text–image similarity (CLIP-T) [18] between the generated image and the background description extracted from the prompt, which provides a complementary measure of prompt alignment. Together, these evaluations reflect whether contextual tokens regain controllability over background attributes, which is crucial for concept dominance mitigation.

Background-neutralized concept similarity (CLIP-I†). Standard CLIP-based image similarity (often referred to as CLIP-I) computes cosine similarity between CLIP image embeddings of a generated image and the original reference images. However, this similarity can be confounded by background correlation: if a generated image reproduces background patterns similar to those in the reference set, the score may be artificially inflated even when the subject identity is not faithfully preserved.

To address this limitation, we propose a background-neutralized CLIP similarity metric, denoted as CLIP-I†. For each generated image, we first synthesize a background image using its background prompt. We then composite the original reference subject onto this generated background to form a background-matched reference image. Finally, we compute CLIP similarity between the generated image and the processed reference image. By controlling for background content, CLIP-I† better isolates subject fidelity and reduces false-positive similarity induced by background leakage.

Table 1: Comparison of personalization methods before and after applying naive augmentation or TOPA.

Method	BRISQUE ↓	CLIP-IQA ↑	CLIP-I† ↑	CLIP-T ↑	FID ↓	Flan-VQA ↑	BLIP-VQA ↑	LLaMA-VQA ↑	BLIP-VQA-Sim ↑
TI	16.634	0.847	0.667	0.446	247	0.608	0.577	0.693	0.634
TI + naive augment	16.941	0.765	0.623	0.425	**213**	**0.688**	0.521	0.688	0.611
TI + TOPA	**14.574**	**0.871**	**0.671**	**0.472**	220	0.654	**0.683**	**0.773**	**0.652**
Custom	**15.147**	0.826	**0.627**	0.506	294	0.711	0.715	0.764	0.649
Custom + naive augment	16.131	0.815	0.608	0.525	279	0.715	0.750	0.806	0.651
Custom + TOPA	15.740	**0.855**	0.613	**0.534**	**259**	**0.717**	**0.810**	**0.844**	**0.668**
DreamBooth LoRA	19.937	0.768	0.717	0.311	181	0.367	0.165	0.273	0.572
DreamBooth LoRA + naive augment	15.502	0.603	0.705	0.332	**160**	0.406	0.185	0.285	0.571
DreamBooth LoRA + TOPA	**15.496**	**0.908**	**0.727**	**0.352**	179	**0.410**	**0.325**	**0.427**	**0.596**
ELITE	15.711	0.832	**0.691**	0.427	219	0.752	0.622	0.915	0.723
ELITE + naive augment	14.699	0.819	0.676	0.440	181	0.775	0.670	0.938	0.719
ELITE + TOPA	**14.244**	**0.858**	0.674	**0.453**	**176**	**0.807**	**0.681**	**0.944**	**0.736**
P+	18.898	0.822	**0.684**	0.430	232	0.583	0.523	0.642	0.616
P+ + naive augment	19.595	0.849	0.671	0.457	**216**	**0.627**	0.620	0.699	0.637
P+ + TOPA	**15.319**	**0.893**	0.676	0.469	217	0.592	**0.683**	**0.746**	**0.637**

4.3 Main Results

We evaluate the effectiveness of TOPA across five representative personalization methods, including Textual Inversion, Custom Diffusion, DreamBooth-LoRA, ELITE, and P+. Tab. 1 summarizes quantitative results in terms of image quality, distribution alignment, background controllability, and subject fidelity.

Across all methods, TOPA maintains or improves perceptual quality compared to original methods. BRISQUE scores improve by up to 22%, for example, TI decreases from 16.634 to 14.574, and CLIP-IQA increases by up to 18%, as in DreamBooth LoRA increasing from 0.768 to 0.908. TOPA consistently reduces FID across all methods, with improvements up to 19.6%, such as ELITE reducing from 219 to 176. This shows that background manipulation and attention rebalancing do not compromise visual realism or alignment with the target concept distribution. TOPA produces a more representative and diverse training set, leading to improved generalization.

VQA-based accuracy and CLIP-T increase with TOPA. For instance, BLIP-VQA accuracy rises from 0.622 to 0.681 for ELITE, and CLIP-T improves from 0.506 to 0.534 for Custom Diffusion. This demonstrates that the two-stage augmentation effectively restores control over background attributes and mitigates concept dominance. CLIP-I† is either preserved or improved. Take TI as an example, CLIP-I† improves from 0.667 to 0.671.

Overall, TOPA consistently improves image quality, distribution alignment, and background controllability while preserving subject fidelity. These results establish TOPA as a general and effective framework for mitigating concept dominance in personalized diffusion models.

4.4 TOPA with Other Context Improving Methods

We evaluate the effectiveness of TOPA on two personalization methods that explicitly aim to improve generation context consistency: Compositional Inversion [30] and Attend-And-Excite [1]. We also evaluate TOPA on Pivotal Tuning

Table 2: Comparison of Compositional Inversion, Attend-And-Excite, and Pivotal Tuning Inversion LoRA with and without TOPA.

Method	BRISQUE↓	CLIP-IQA↑	CLIP-I†↑	CLIP-T↑	FID↓	Flan-VQA↑	BLIP-VQA↑	LLaMA-VQA↑	BLIP-VQA-Sim↑
Compositional Inversion	16.620	0.883	0.642	**0.471**	279	**0.627**	0.623	0.717	0.640
Compositional Inversion + TOPA	**15.971**	**0.888**	**0.658**	0.454	**225**	0.625	**0.646**	**0.722**	**0.645**
Attend-And-Excite	**16.421**	0.837	**0.594**	0.485	296	0.618	0.567	0.744	0.649
Attend-And-Excite + TOPA	18.568	**0.846**	0.557	**0.512**	**256**	**0.635**	**0.703**	**0.796**	**0.668**
Pivotal Tuning Inversion LoRA	**16.170**	0.816	0.715	0.392	187	0.513	0.415	0.516	0.602
Pivotal Tuning Inversion LoRA + TOPA	16.609	**0.840**	**0.743**	**0.430**	**147**	**0.577**	**0.578**	**0.642**	**0.636**

Table 3: Textual Inversion ablation results with background replacement or different loss augmentations. Best metrics are bolded, and second-best metrics are underlined.

Method	BRISQUE↓	CLIP-IQA↑	CLIP-I†↑	CLIP-T↑	FID↓	Flan-VQA↑	BLIP-VQA↑	LLaMA-VQA↑	BLIP-VQA-Sim↑
Original	16.634	0.847	0.667	0.446	247	0.608	0.577	0.693	0.634
With background replaced only	16.899	<u>0.871</u>	0.661	**0.484**	<u>231</u>	0.610	<u>0.673</u>	0.731	**0.654**
With $\mathcal{L}_{\text{ctx}}$ only	<u>15.346</u>	0.855	<u>0.669</u>	0.453	251	0.608	0.608	0.727	0.634
With $\mathcal{L}_{R^*}$ only	17.166	0.846	0.650	0.470	269	<u>0.650</u>	0.640	<u>0.740</u>	0.643
With both losses	**14.574**	**0.871**	**0.671**	<u>0.472</u>	**220**	**0.654**	**0.683**	**0.773**	<u>0.652</u>

Inversion LoRA [20], which masks the personalization loss in latent space to focus optimization on the subject region. Tab. 2 summarizes quantitative results for these methods with and without TOPA.

Across these methods, applying TOPA consistently improves key metrics. For Compositional Inversion, BRISQUE improves from 16.620 to 15.971, indicating better perceptual quality, while CLIP-IQA and CLIP-I† increase, demonstrating improved subject fidelity. Background controllability measured by CLIP-T remains stable, and FID drops from 279 to 225, showing enhanced distribution alignment. VQA-based metrics for prompt adherence also improve or remain high, reflecting stronger background and subject consistency.

Similarly, Attend-And-Excite benefits from TOPA. Although BRISQUE slightly increases, background controllability improves as CLIP-T rises from 0.485 to 0.512. FID decreases from 296 to 256, indicating a better match to the target distribution. All VQA metrics increase, confirming that TOPA enhances prompt-driven generation while preserving subject identity.

For Pivotal Tuning Inversion LoRA, TOPA improves subject fidelity and prompt alignment. CLIP-IQA increases from 0.816 to 0.840, CLIP-I† from 0.715 to 0.743, and CLIP-T from 0.392 to 0.430. FID decreases from 187 to 147, and all VQA-based metrics improve, indicating stronger prompt adherence and context consistency while preserving the personalized subject.

Overall, these results indicate that TOPA effectively improves generation quality for these methods, enhancing perceptual quality, distribution alignment, and prompt adherence simultaneously.

4.5 Ablation Study

Naive Augmentation Alone Fails We first evaluate standard augmentation, applying random crop, flip, color jitter, blur, and noise 100 times per reference

image without using TOPA. Tab. 1 shows that naive augmentation yields inconsistent results. For instance, Flan-VQA accuracy for TI increases 13% from 0.608 to 0.688, but CLIP-I† drops 6.6% from 0.667 to 0.623. BRISQUE also worsens for Custom Diffusion, increasing from 15.147 to 16.131.

In contrast, applying TOPA consistently improves performance across all metrics. For TI, Flan-VQA rises to 0.654, CLIP-I† to 0.671, BRISQUE improves to 14.574. Similarly, Custom Diffusion and other methods show clear improvements with TOPA, demonstrating that naive augmentation alone is insufficient and that TOPA is necessary for robust gains in image quality, distribution alignment, and subject-background fidelity.

Both Losses Are Needed We examine the effectiveness of contextual loss $\mathcal{L}_{\text{ctx}}$ and perturbation loss $\mathcal{L}_{R^*}$. Tab. 3 shows that $\mathcal{L}_{\text{ctx}}$ slightly improves perceptual quality and subject fidelity by improving BRISQUE by 7.8%. $\mathcal{L}_{R^*}$ improves background metrics slightly by increasing CLIP-T by 5.2% and BLIP-VQA by 6.3%. However, BRISQUE increases, indicating worse image quality.

Compared to using one loss only, combining both losses achieves the best results. BRISQUE improves 12.5%, CLIP-IQA increases 2.8%, FID improves 10.8%, and BLIP-VQA increases 10.6%. Both objectives are necessary for effective subject and background disentanglement.

Background Replacement Alone Is Insufficient We further investigate whether replacing the backgrounds of the dataset without perturbations can work. Tab. 3 compares original TI, TI with background replacement, and TI with TOPA.

Replacing backgrounds alone improves several background-related metrics. CLIP-T increases 8.5%, and BLIP-VQA accuracy increases 16.6%. FID also decreases 6.5%, suggesting improved distribution alignment. However, subject fidelity slightly degrades, and perceptual quality does not improve.

Compared to merely replacing backgrounds, TOPA achieves stronger overall performance. BRISQUE improves 12.4%, and FID improves 10.9%. VQA metrics also improve consistently, with BLIP-VQA accuracy increasing 18.4%.

These results show that background replacement alone cannot fully address concept dominance. Perturbation optimization is necessary to maintain subject fidelity while improving background controllability.

4.6 Qualitative Evaluation

We present qualitative results generated with TI in Fig. 4. Additional qualitative results for other personalization methods are provided in Appendix.

As shown in the figure, our method successfully includes the generation context in the generated images while preserving the personalized concept. For example, in the upper-right case, the original TI result produces a mountain background instead of the requested city, whereas TOPA clearly introduces the city context. Similarly, in the lower-left case, TOPA not only generates the purple appearance but also produces a shape that more closely resembles the reference concept. These examples demonstrate that TOPA improves contextual controllability while maintaining concept fidelity in terms of human view.

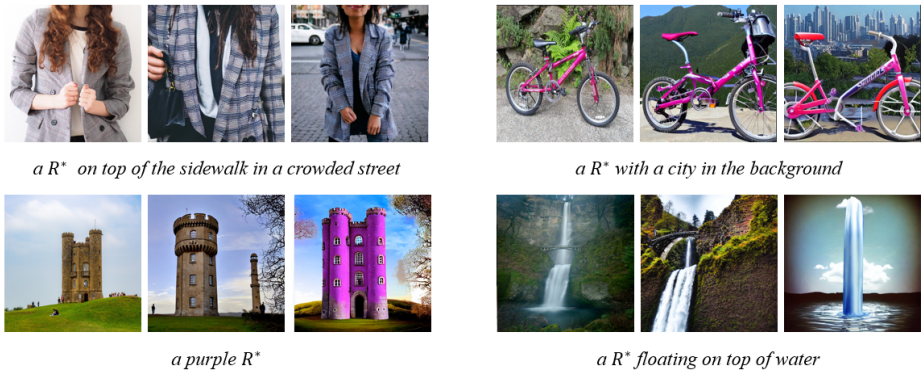

Fig. 4: Examples of concept-dominance failures and the effect of TOPA. For each example, the left column shows the reference images. The middle column shows generations from **TI** *without* TOPA, where the learned concept dominates and suppresses prompt-specified context. The right column shows generations *with* TOPA, which restores context controllability and improves compositionality. R^* in the prompts represents the learned concept token.

5 Conclusion

We proposed **Target-Oriented Perturbation-based Augmentation (TOPA)**, a data augmentation pipeline that improves concept–context compositionality without modifying personalization architectures, training objectives, or inference procedures, making it compatible with a wide range of existing personalization pipelines. TOPA introduces two complementary components: **Cross-Attention Rebalancing Module (CARM)** to guide token–region alignment during personalization, and **Subject-Isolation Compositing (SIC)** to increase background diversity. Extensive experiments across multiple diffusion personalization methods demonstrate that TOPA consistently mitigates concept dominance and generally improves image quality, while preserving subject fidelity. We hope our work can encourage further research on improving personalization by augmenting datasets.

Acknowledgment

This research is supported in part by the National Research Foundation, Prime Minister's Office, Singapore, and the Ministry of Digital Development and Information, under its Online Trust and Safety (OTS) Research Programme (MDDI-OTS-001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Prime Minister's Office, Singapore, the Ministry of Digital Development and Information, or the Centre for Advanced Technologies in Online Safety. This research is supported in part by the Energy Research Institute @ NTU, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore.

References

1. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
2. Chen, W., Song, D., Li, B.: Trojdiff: Trojan attacks on diffusion models with diverse targets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4035–4044 (2023)
3. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. In: CVPR (2019)
4. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: CVPR Workshops (2020)
5. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
6. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. Advances in Neural Information Processing Systems **36**, 15890–15902 (2023)
7. Harris, E., Marcu, A., Painter, M., Niranjan, M., Hare, J.: Fmix: Enhancing mixed sample data augmentation. In: ArXiv preprint arXiv:2002.12047 (2020)
8. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: P+: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
10. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. arXiv preprint arXiv:2212.04488 (2022)
11. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36**, 30146–30166 (2023)
12. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
13. Liang, C., Wu, X.: Mist: Towards improved adversarial examples for diffusion models. arXiv preprint arXiv:2305.12683 (2023)
14. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR). pp. 723–727. IEEE (2011)
15. Motamed, S., Paudel, D.P., Van Gool, L.: Lego: Learning to disentangle and invert personalized concepts beyond object appearance in text-to-image diffusion models. arXiv preprint arXiv:2311.13833 (2023)
16. Müller, S., Hutter, F.: Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In: ICCV (2021)
17. Park, S., Choi, S., Park, H., Yun, S.: Steering guidance for personalized text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15907–15916 (2025)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
19. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242 (2022)
20. Ryu, S.: Low-rank adaptation for fast text-to-image diffusion fine-tuning, `https://github.com/cloneofsimo/lora`
21. Styborski, J., Lyu, M., Lu, J., Kapur, N., Kong, A.W.K.: When and where do data poisons attack textual inversion? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19439–19449 (2025)
22. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
23. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
24. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
25. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
26. Wu, F., Pang, Y., Zhang, J., Pang, L., Yin, J., Zhao, B., Li, Q., Mao, X.: Core: Context-regularized text embedding learning for text-to-image personalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8377–8385 (2025)
27. Xue, H., Liang, C., Wu, X., Chen, Y.: Toward effective protection against diffusion-based mimicry through score distillation. In: The Twelfth International Conference on Learning Representations (2023)
28. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)
29. Zhang, H., Cisse, M., Dauphin, Y., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
30. Zhang, X., Wei, X.Y., Wu, J., Zhang, T., Zhang, Z., Lei, Z., Li, Q.: Compositional inversion for stable diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7350–7358 (2024)