
AN h -SPACE BASED ADVERSARIAL ATTACK FOR PROTECTION AGAINST FEW-SHOT PERSONALIZATION

Xide Xu

Computer Vision Center
Universitat Autònoma de Barcelona
Barcelona, Spain
xide@cvc.uab.cat

Sandesh Kamath

Computer Vision Center
Universitat Autònoma de Barcelona
Barcelona, Spain
skamath@cvc.uab.cat

Muhammad Atif Butt

Computer Vision Center
Universitat Autònoma de Barcelona
Barcelona, Spain
mabutt@cvc.uab.cat

Bogdan Raducanu

Universitat Autònoma de Barcelona
Barcelona, Spain
bogdan@cvc.uab.cat

ABSTRACT

The versatility of diffusion models in generating customized images from few samples raises significant privacy concerns, particularly regarding unauthorized modifications of private content. This concerning issue has renewed the efforts in developing protection mechanisms based on adversarial attacks, which generate effective perturbations to poison diffusion models. Our work is motivated by the observation that these models exhibit a high degree of abstraction within their semantic latent space (termed ' h -space'), which encodes critical high-level features for generating coherent and meaningful content. In this paper, we propose a novel anti-customization approach, called HAAD (h -space based Adversarial Attack for Diffusion models) that leverages adversarial attacks to craft perturbations based on the h -space that can efficiently degrade the image generation process. Building upon HAAD, we further introduce a more efficient variant, HAAD-KV, that constructs perturbations solely based on the KV parameters of the h -space. This strategy offers a stronger protection, that is computationally less expensive. Despite their simplicity, our methods outperform state-of-the-art adversarial attacks, highlighting their effectiveness.

[Warning: This paper may contain images that could produce visual discomfort.]

Keywords Diffusion model, Adversarial attack, New concept learning, Text-to-image generation, Privacy Protection

1 Introduction

Diffusion Models [7, 8, 9] are the current state of the art for image generation, outperforming GANs in terms of image quality and mode coverage [10, 11]. With the introduction of large-scale diffusion models, [14, 12, 13] able to generate intricate details and complex patterns via textual prompts, we achieve unmatched accuracy and robustness towards generating high-fidelity images. Image editing [20, 21], image-to-image translation [22], text-to-3D images synthesis [23, 24, 25], video generation [26, 27, 28], anomaly detection in medical images [29], etc. are few of the many applications of diffusion models. One application that has become extremely popular due to its versatility and ease of use is the customization of diffusion models with personal content [48, 47, 46]. These models have the ability to create personalized content from few user images by fine-tuning a pre-trained diffusion model (e.g. Stable Diffusion) in order to learn how to bind a unique token to a novel concept. Consequently, we can generate novel views in different contexts and visualize them under different artistic styles.

Few-shot image customization leverages diffusion models to create user-tailored content, adapting the output to fit particular preferences. This opens new ways of creating a more engaging and meaningful interaction with AI systems. The ability of diffusion models to learn and adapt to specific user inputs has revolutionized the field of personalized

content creation, making it more accessible and impactful across various industries, such as visual arts (artworks), virtual reality, gaming, and e-commerce. While these customization approaches are powerful tools for generating user-specific content, they also represent significant privacy risks. Malicious individuals could exploit the vulnerability of this technology to produce and spread deceptive images ('deep fakes') that are visually indistinguishable from genuine ones [30]. The negative effects of privacy risks induced by deep fakes could span from information leakage, unauthorized reproduction of artwork [31, 32] to extreme cases where it could severely impact an individual's personal life and reputation [33].

In order to prevent these privacy risks, there are currently some efforts that explore protection mechanisms that can prevent against the malicious use of customized diffusion models [41]. The main idea of the anti-customization methods is to obtain a protected image i.e an adversarial example, from a text-to-image (T2I) diffusion model, with the well-known Projected Gradient Descent (PGD) algorithm [56]. This protected image when used by an attacker to train a diffusion model, leads to its poisoning, thereby successfully preventing it from generating good images. For a pictorial representation of this flow, refer to Figure 1. Recently, both targeted and untargeted approaches have been proposed for attacking the T2I diffusion models with adversarial examples [62, 60]. Targeted attacks are designed to disrupt a model's functionality by forcing it to produce a specific, predetermined output (e.g. by altering a generated image to match a specific pattern) [82]. On the other hand, untargeted attacks [53, 63, 65] disrupt the overall model's functionality, without guiding it towards a specific output. One of the limitations of targeted attacks is that they may leave visible traces of the specific pattern in the protected image, which a skillful person could potentially analyze in order to purify and recover the original output.

This remarkable capability of diffusion models to generate user-specific content is rooted in their underlying semantic latent space [34], termed ' h -space', which is responsible for generating high-quality images, coherent with the user's input. h -space represents the deep features from the middle-block of the U-Net architecture within the denoiser component of the diffusion model (see Fig. 2). Leveraging this high degree of abstraction within the h -space, we propose an adversarial attack which serves as an anti-customization method, named HAAD (**h -space based Adversarial Attack for Diffusion models**). The imperceptible perturbations introduced by our approach disrupts the model's ability to maintain consistency with user-specific input, effectively interfering with the generation of personalized content and hence, providing a strong protection to the personal data. Building on prior works [46, 16] that emphasize the critical role of cross-attention mechanisms in guiding semantic alignment and content fidelity within diffusion models, we further construct a more efficient variant of our method, termed HAAD-KV. Instead of considering the entire h -space, HAAD-KV focuses solely on disrupting the key (K) and value (V) parameters within the cross-attention layer of the U-Net's h -space. This targeted intervention leverages the high semantic influence of the attention pathway while requiring significantly fewer parameter ($\sim 0.05\%$) updates w.r.t. HAAD. As a result, HAAD-KV achieves enhanced attack performance with reduced computational overhead.

We summarize our contributions in the list below:

- we propose HAAD, a simple yet effective and robust approach for privacy protection by constructing a perturbation based on the ' h -space' of diffusion models.
- additionally, we introduce a more efficient variant, by focusing only on the KV parameters of the cross-attention layer (HAAD-KV), which provides increased protection at a small fraction of computational cost.
- we show through extensive comparison with several models and validation datasets that our approach and its variant not only outperform state-of-the-art methods based on adversarial attack, but also present increased robustness against a variety of purification strategies.

2 Related Work

2.1 Personalized Diffusion Models

The personalization of text-to-image (T2I) diffusion models has gained significant attention due to their ability to generate user-specific content. This process typically requires only a few reference images (commonly 3–5, or around 20 for high-quality face images) and trains a model to associate a unique identifier with the target concept, bridging the gap between generic AI-generated imagery and highly customized visual content. Early approaches to personalization focused on text embedding adaptation rather than model fine-tuning. Textual Inversion [48] pioneered this direction by learning a compact text embedding vector to represent a new concept through inverting visual examples into the pretrained model's text space. While parameter-efficient, this method relies heavily on the pretrained knowledge and struggles to capture complex visual details compared to fine-tuning. By better fine-tuning the model, DreamBooth [47] achieves improved personalized generation. While effective, this approach is computationally expensive and risks

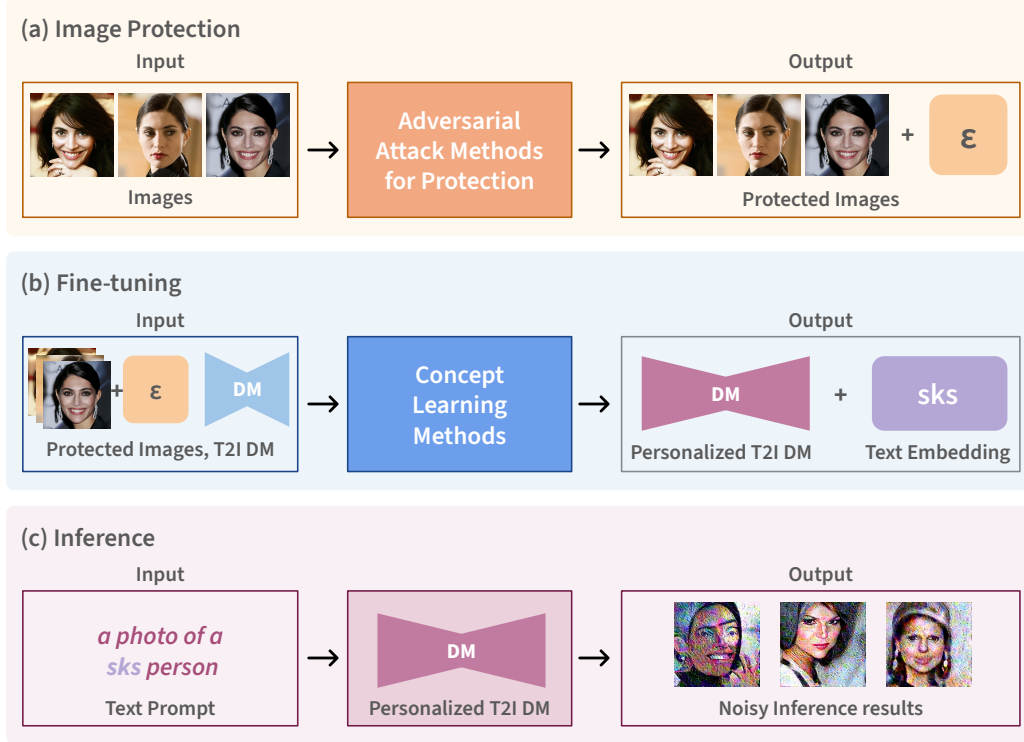


Figure 1: A step-by-step explanation of protection against customization by few-shot personalization methods using adversarial perturbations: a) Starting with a set of clean images, these methods obtain the protected images using adversarial attack methods (e.g. PGD); b) During fine-tuning, a diffusion model is personalized with these protected images leading to a poisoned model; c) at inference, the poisoned model will not be able to generate good customized images.

overfitting. To mitigate this, Custom Diffusion (CD) [46] introduces and fine-tunes additional cross-attention layers while keeping the original model frozen. This selective adaptation significantly reduces computational costs while preserving the model’s generalization, offering a more efficient and scalable solution for personalization. Another method which provides an efficient way for parameter fine-tuning is SVDiff [49]. SVDiff uses Singular Value Decomposition (SVD) to create a low-rank approximation of weight updates potentially offering even greater parameter efficiency. More recent approaches include [50, 51, 52].

2.2 Anti-Customization of Diffusion Models

A number of recent works have explored adversarial attacks as a protection method to counter unauthorized customization using diffusion models. Among the earliest, AdvDM [60] introduces the idea of generating adversarial noise during training by maximizing the model’s original loss function, thereby preventing effective learning from perturbed inputs. While effective in scenarios such as artwork protection, its optimization focuses on the training loss rather than internal semantics, making the protection susceptible to adaptation by personalized models.

Other approaches, such as PhotoGuard [62], Mist [61], and ACE [82], generate visually imperceptible noise using fixed global patterns—ranging from random noise to high-frequency Moiré textures—to guide model outputs toward predefined targets or degradation. However, these patterns operate at the pixel level, lacking alignment with the model’s internal representations. As a result, they often fail to disrupt semantics consistently in the generated content. Worse yet, they may leave visually detectable patterns.

Several methods attempt to improve the generality and robustness of the protection by introducing perturbations at the training stage. Anti-DreamBooth [53] and its extension MetaCloak [54] inject perturbations using surrogate models, sometimes combining loss components to induce instability in the model’s optimization. However, these approaches still do not leverage the semantic abstraction capacity of diffusion models, instead relying on large-scale or iterative optimization that is costly and often model-specific.

CAAT [63] takes a more parameter-efficient route by targeting only the cross-attention layers in the customized diffusion (CD) model, noting that these components undergo the most change during personalization. While this improves efficiency, it treats the attention modules as a monolithic block and does not distinguish between different attention projections.

In this work, we overcome both semantic and efficiency limitations through two key innovations. First, we propose HAAD, which injects perturbations into the h -space, which leads to stronger misalignment between input identity and output semantics, even under small noise budgets. Second, we introduce a more efficient variant, HAAD-KV, that perturbs only the KV parameters of the cross-attention layer within the h -space. Together, these innovations result in a stronger, more resilient protection that outperforms prior state-of-the-art approaches in both effectiveness and computational efficiency.

3 Method

3.1 Latent Diffusion Models

A diffusion model consists of a forward process where the input image is disrupted by adding noise in multiple steps and a reverse (i.e. generative) process where the final image is obtained by applying multiple denoising steps. A latent diffusion model (LDM)[14] is a diffusion model where the diffusion processes are applied to the latent space instead of the image space. LDM consists of two main components: (i) an autoencoder (\mathcal{E}) that transforms an image x into a latent code $z_0 = \mathcal{E}(x)$, while the decoder (\mathcal{D}) reconstructs the latent code back to the original image such that $\mathcal{D}(\mathcal{E}(x)) \approx x$; and (ii) a diffusion model (parameterized by θ), which applies the diffusion processes on the latent space, commonly a U-Net [71] based model which can be conditioned using class labels, segmentation masks, or textual input. Let $\tau_\theta(y)$ represent the conditioning mechanism (e.g. prompt) that converts a condition y into a conditioning vector and $t \in T$ be the number of steps of the diffusion process. The reconstruction loss used to train the model is :

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0, y, \epsilon \sim \mathcal{N}(0,1)} \|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2, \quad (1)$$

where, ϵ_θ is the conditional U-Net [71] that predicts the noise added in the denoising step.

3.2 Adversarial Attacks on Diffusion Models

Adversarial attacks compute a subtle, human-imperceptible perturbation, added to the input data which gets grossly misclassified. It exposes the brittleness of deep learning classification models which obtain super-human performance. An adversarial attack is formulated as follows: given an input x , obtain a perturbed input x' such that :

$$\underset{x'}{\operatorname{argmax}} \mathcal{L}_\phi(x') \quad \text{s.t. } \|x' - x\|_\infty \leq \eta \quad (2)$$

where, \mathcal{L}_ϕ is the classification loss function i.e. cross-entropy used to train the model with parameters ϕ . We use the ℓ_∞ -norm to control the noise budget given by η . Commonly, the strong iterative PGD algorithm [56] is used to construct the adversarial attack. For diffusion models, the main focus of this work, the above objective remains the same: obtaining an adversarial image x' from a clean image x perturbed with an imperceptible noise (within η -budget).

3.3 HAAD: h -space based Adversarial Attack on Diffusion Models

The remarkable capability of diffusion models to generate user-specific content is rooted in their underlying semantic latent space [34]. This semantic latent space (' h -space'), represents the deep features from the middle-block of the U-Net architecture (see Fig. 2). Kwon et al [34] found that the h -space exhibits nice properties like homogeneity, linearity, robustness, and consistency across timesteps, similar to the latent space in GANs. By exploring the properties of this space, researchers have been able to discover meaningful semantic directions that have application to image editing (manipulating facial appearance) [36, 37, 39, 38] and image-to-image translation [35].

Inspired by the strong properties of the h -space, we devise an adversarial attack leveraging this rich semantic feature (illustrated in Figure 2). We use the gradient of the reconstruction loss (\mathcal{L}_{LDM}) restricted to the h -space to construct the adversarial perturbation using the iterative PGD method.

Let W_h represent the features (weights) of the h -space, and δ denote the adversarial perturbation added to disrupt the semantic structure of the h -space. The perturbation is computed iteratively using the Projected Gradient Descent(PGD)

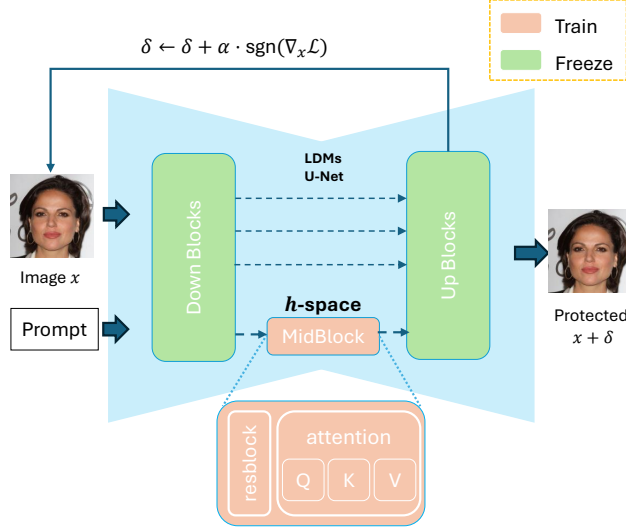


Figure 2: Block diagram of HAAD method constructing the perturbed image $x + \delta$ based on the gradient of the loss function wrt to the h -space. Only h -space is optimized during training, while the rest of the model is kept frozen.

method approximately formulated as:

$$x^{t+1} = \Pi_{\eta} \left(x^t + \alpha \cdot \text{sign} \left(\nabla_{x^t} \mathcal{L}_{\text{LDM}}(x^t, W_h^t) \right) \right),$$

where: \mathcal{L}_{LDM} is the reconstruction loss (Eq 1), $\Pi_{\eta}(\cdot)$ is the projection operator ensuring that $\|\delta\|_{\infty} \leq \eta$, where η is the noise budget, α is the step size for the gradient update, t represents the current iteration. W_h^t indicates that the gradients were calculated with the updated h -space after t step.

Following, we explain the reasoning behind training the model simultaneously while constructing the attack. Adversarial attacks can be constructed either to the full model or selectively to specific components. Attacks that are integrated into the training loop—such as those embedded within fine-tuning processes—interact with the model’s learning dynamics. These methods can adaptively influence the model’s parameter updates, leading to more effective adversarial perturbations. The same strategy was also employed by CAAT. In HAAD, we only update the parameters corresponding to the h -space using the default optimization setting for the Stable Diffusion. By restricting the update to the h -space, we ensure that the learned semantic features are directly affected by the adversarial perturbation while keeping the rest of the model intact. This tightly integrated training strategy enables the perturbation to interfere more effectively with the internal learning process of diffusion models—resulting in robust, semantically aligned protection that are highly transferable across personalization settings.

During training, we jointly perform two complementary operations within each optimization step. First, we compute the loss \mathcal{L}_{LDM} used in latent diffusion models, and then apply a PGD step to determine a noise vector that maximizes the disruption of the reconstruction objective when injected into the image. This perturbation is added to the training images before the next iteration, encouraging the model to learn under adverse semantic conditions. This is repeated operation leads to a stronger protection.

Algorithm 1, presents the pseudocode of our untargeted adversarial attack to generate perturbations based on the h -space as protection against unauthorized customization by a diffusion model. In our implementation, we freeze the entire diffusion model, only keeping the h -space trainable and iteratively update its weights (similar to CAAT[63]), while constructing the attack perturbation to obtain strong protection.

HAAD-KV: An Improved Variant of HAAD. While HAAD introduces adversarial noise into the h -space to disrupt high-level semantic representations, we further refine this strategy by narrowing the scope of perturbation to the key (K) and value (V) parameters within the cross-attention layers in the h -space. This variant, referred to as HAAD-KV, leverages the fact that cross-attention plays a pivotal role in text-to-image diffusion models by creating an alignment between the input prompt and the generated visual content.

The motivation behind HAAD-KV stems from two key observations. First, during the personalization process, the cross-attention layers—especially the KV parameters—undergo substantial adaptation to capture new concepts. These components control how visual features attend to the semantic prompt over time, making them critical to preserving

Algorithm 1 HAAD: h -space based Adversarial Attack

Input: Image x , h -space parameters W_h , step length α , epochs N , budget η , learning rate l

Output: Perturbed image x'

```
1: Initialize  $\delta$ 
2: for  $n = 1 \rightarrow N$  do
3:    $\nabla_{W_h}, \nabla_x \leftarrow \nabla_{LDM}(W_h, x + \delta)$ 
4:    $W_h \leftarrow W_h - l \nabla_{W_h} \triangleright h$ -space weight updated.
5:    $\delta \leftarrow \delta + \alpha \text{sgn}(\nabla_x) \triangleright$  perturbation  $\delta$  updated.
6:   if  $\delta > \eta$  then
7:      $\delta \leftarrow \text{clip}(\delta, -\eta, \eta) \triangleright$  ensures  $\|\delta\|_\infty \leq \eta$ 
8:   end if
9:    $x' \leftarrow x + \delta$ 
10: end for
```



Figure 3: Images generated by popular diffusion models for customization: LoRA+Dreambooth (first two rows), Custom Diffusion (bottom two rows), using different adversarial protection methods with a noise budget of 4/255.

identity and fidelity in customized generation. Second, by focusing the attack on these parameters, we are able to maximize semantic disruption while updating only a minimal subset of parameters, thereby improving computational efficiency and reducing the perceptual footprint of the perturbation.

In practice, HAAD-KV operates similarly to HAAD in terms of optimization: we retain the PGD-based perturbation strategy guided by the reconstruction loss \mathcal{L}_{LDM} , but constrain the perturbation injection and gradient updates only to the KV parameters of the cross-attention layer within the h -space. All other parts of the model remain frozen.

This approach not only enhances the interpretability of the attack — by isolating the exact mechanism through which text-image alignment is corrupted—but also leads to stronger overall attack performance. As demonstrated in our experiments (Section 4), HAAD-KV consistently achieves greater degradation of personalized content than HAAD, despite requiring fewer parameter updates. These results highlight HAAD-KV as a highly effective and efficient protection strategy against unauthorized diffusion model customization. In Supp. Material, Sec. 6.7, we present a theoretical framework and supporting analysis that illustrates the relationship within the h -space - specifically the key and value (KV) parameters in cross-attention layers - and the model’s semantic structure. We show that targeting these components can lead to significant semantic misalignment, thereby enhancing the strength of the protection.

Table 1: Comparison of our methods with other adversarial attack methods on LoRA+DB and CD. **bold** indicates the best, while underline indicates the second best. HAAD-KV achieves the best performance in 10 out of 12 metric-dataset combinations

	CelebA-HQ								WikiArt			
	LoRA+DB				CD				LoRA+DB		CD	
	CI \uparrow	CS \downarrow	FDR \uparrow	ISM \downarrow	CI \uparrow	CS \downarrow	FDR \uparrow	ISM \downarrow	CI \uparrow	CS \downarrow	CI \uparrow	CS \downarrow
No Attack	21.32	83.13	0.005	0.6904	20.89	85.64	0.005	0.6907	34.62	64.51	34.22	66.36
AdvDM	24.02	76.79	0.005	0.7633	23.87	77.23	0.015	0.6721	35.32	61.12	36.45	61.85
ACE	28.22	72.26	0.015	0.6081	26.65	73.47	0.055	0.6065	37.01	58.69	49.12	59.84
ACE+	28.67	73.14	0.000	0.6097	28.20	74.01	0.045	0.6069	37.11	59.16	52.90	60.13
CAAT	<u>30.96</u>	72.17	0.080	0.5547	<u>29.31</u>	<u>73.26</u>	0.185	0.5763	37.52	58.64	51.17	<u>59.73</u>
HAAD	29.10	72.06	0.085	0.5175	29.10	73.35	0.150	0.5657	37.66	58.59	52.14	59.78
HAAD-KV	31.82	71.91	0.100	0.5083	29.52	72.98	0.185	0.5606	37.88	58.26	<u>52.86</u>	59.52

4 Experimental Results

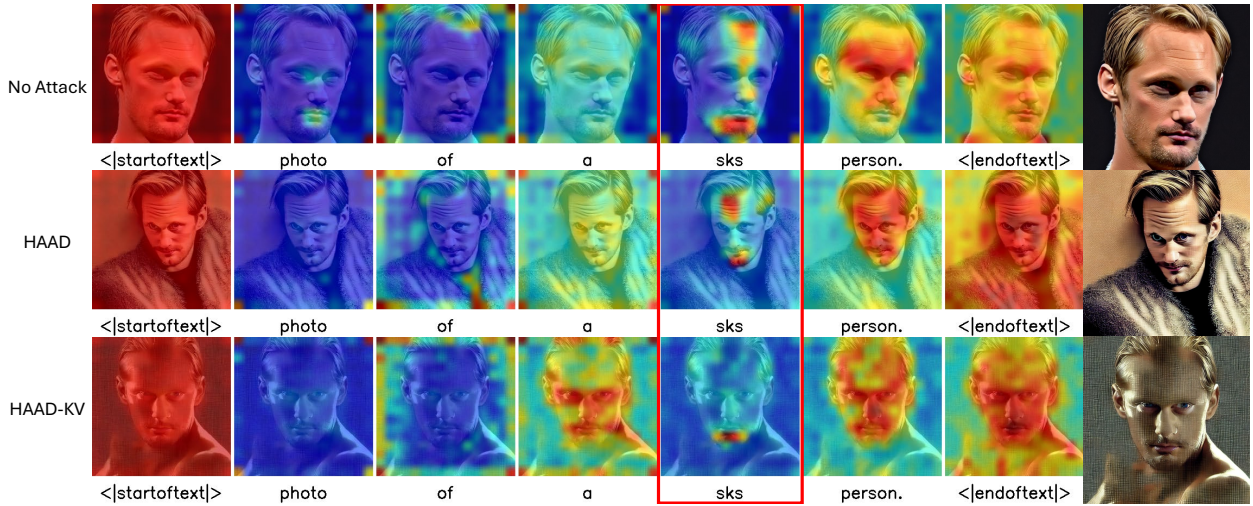


Figure 4: Visualization of the cross attention map of each token in the h -space block during inference time: No Attack (first row), HAAD (middle row), and HAAD-KV (bottom row).

4.1 Experimental Setup

Datasets. For the validation of our methods, we conducted experiments on two widely used datasets: CelebA-HQ [69] and WikiArt [83] datasets. For CelebA-HQ, we select 200 images, corresponding to 10 persons (20 images per person), ensuring diversity in terms of gender, ethnicity and age. Similarly, for WikiArt, we choose 200 paintings from 10 artists (20 paintings per artist).

Customization Methods. We evaluate the effectiveness of our protection approach by applying it to several widely used customization pipelines. Specifically, we test on LoRA+DreamBooth [81, 47] (abbreviated as LoRA+DB) and Custom Diffusion (CD) [46], both of which represent state-of-the-art few-shot personalized concept learning approaches. These models capture different paradigms of personalization: LoRA+DB relies on parameter-efficient adaptation, while CD focuses on fine-tuning only the additional attention layers—providing a comprehensive evaluation ground for our methods. To ensure the reproducibility of our results, we use the publicly available code repositories for each method and follow their official hyper-parameter settings throughout our experiments.

Comparison with SotA. We compare the performance of HAAD and HAAD-KV with several representative adversarial based protection approaches, including AdvDM [60], ACE, ACE+ [82], and CAAT [63]. These methods have been widely used in recent studies on protecting content from unauthorized diffusion model customization and provide a solid baseline for evaluating the effectiveness and efficiency of our proposed attacks.

Details of the Adversarial Attacks. Our methods were computed based on SD1.5 as the reference model, as it is the most commonly deployed diffusion model. HAAD training focuses exclusively on the h -space of Diffusion model [15], using a batch size of 1, the learning rate of 1×10^{-5} for 250 training steps. Mixed precision of bf16 is employed. The conditioning prompt used was "photo of a sks person/painting", where the 'sks' user-specific token is initialized with 'ktn'. For the PGD attack, we set $\alpha = 5 \times 10^{-3}$ (step size) and a noise budget of $\eta = 4/255$ in ℓ_∞ -norm. All the experiments were run on a server with NVidia A40 GPU. For reproducibility, we provide all the settings in Supp. Material, Tables 7 and 8.

Adversarial perturbation based anti-customization methods have studied their impact using different noise budgets (η). AdvDM was among the first works to report with a larger noise budget of $\eta = 32/255$ and similarly, CAAT with 0.1 ($\approx 25.5/255$). Larger budgets tend to introduce perceptible noise that a human can easily detect visually, making them less suitable for real-world applications. Recently, ACE used a challenging setting with the noise budget of $\eta = 4/255$, where the noise in the perturbed image is indeed imperceptible. We choose the same strict setting of $\eta = 4/255$ in all our experiments to show that our approach works in the strictest condition. However, we perform a study to show the results of our approach with varying noise budgets (refer to Supp. Material, Figure 5 and Table 9).

Evaluation metrics. To assess the effectiveness of the evaluated protection methods, we adopted several quantitative metrics. CLIP Image-to-Image Similarity (CLIP-SIM) [73] (shortened CS in tables), measures semantic similarity between generated and reference images based on CLIP embeddings. Lower scores indicate stronger protection and semantic misalignment. We additionally use CLIP-IQA [73] (shortened CI in tables), which evaluates perceptual image quality from a CLIP perspective. Higher CI scores reflect lower image quality and thus stronger disruption. For CelebA-HQ, we include two face-specific metrics: Face Detection Failure Rate (FDFR) [75], where higher values indicate successful degradation of facial content, and Identity Score Matching (ISM) [76], where lower values indicate better obfuscation of personal identity.

4.2 Qualitative Evaluation

Figure 3 shows that our attack significantly outperforms other approaches across LoRA+DB and CD. For LoRA+DB, our method introduces visible alterations both at the structural and semantic levels. For instance, we observe changes in facial attributes such as hair color, and in the case of art images, significant distortion of the main subject along with noticeable shifts in artistic style. For CD, the facial outputs undergo even more pronounced stylistic transformations, deviating substantially from the identity of the source individual. In artwork cases, the protected input leads to generation of entirely different visual compositions, often replacing the original subject with incoherent or unrelated patterns. Additional visual comparisons for each model and content type can be found in Supp. Material, Figures 10, 11, 13, and 14.

4.3 Quantitative Evaluation

Table 1 summarizes the quantitative comparison between our methods (HAAD and HAAD-KV) and several state-of-the-art adversarial based protection techniques across two personalization settings: LoRA+DB and Custom Diffusion (CD). We evaluate performance using a range of perceptual, structural, and semantic metrics across both CelebA-HQ and WikiArt datasets.

We observe that HAAD achieves competitive performance, often outperforming AdvDM, ACE, and ACE+, and performing comparably to CAAT in several settings. Notably, HAAD achieves second-best results in most metrics and occasionally surpasses CAAT, particularly on CelebA-HQ in terms of ISM and FDFR.

However, once the attack is restricted to only the KV parameters in the cross-attention layer, i.e. the HAAD-KV variant, our method consistently achieves the best performance across nearly all metrics and setups. On the CelebA-HQ dataset, HAAD-KV achieves the best performance across all four metrics: it attains the highest CLIP-IQA (CI) and Face Detection Failure Rate (FDFR), as well as the lowest Identity Score Matching (ISM) and CLIP-SIM (CS), demonstrating its strong ability to degrade image quality, obscure identity, and disrupt semantic consistency. On the WikiArt dataset, HAAD-KV consistently ranks first or second in both CI and CS, confirming its effectiveness in protecting artistic content from unauthorized replication. When compared to CAAT, HAAD-KV achieves lower semantic similarity (CS), better perceptual degradation (CI), and more substantial structural distortion, indicating a clear advantage in both performance and efficiency.

These findings highlight the strength of our approach in disrupting both perceptual and semantic fidelity. By consistently outperforming existing methods across diverse models and datasets, HAAD-KV demonstrates its practical effectiveness as a generalizable protection against unauthorized customization.

User study: We conducted an user study to evaluate how perceptually imperceptible the adversarial perturbations added to the images will be when used as a protection mechanism. The results are reported in Supp. Material, Sec. 6.8.

4.4 Protection Explainability and Parameter Efficiency

To better understand the mechanism behind the effectiveness of our approach, we conduct a comparative analysis between HAAD and HAAD-KV from two complementary perspectives: semantic disruption in attention maps and training parameter efficiency.

We begin by visualizing the cross-attention maps corresponding to in the h -space during inference (Figure 4). We compare three settings: clean images (No attack), protected images with HAAD, and protected images with HAAD-KV. Red color highlights areas of the image which shows strong connection between text token and the generated image, while the blue color indicates the absence of such connection.

In the ‘No attack’ case, attention is tightly focused on meaningful facial regions (e.g., forehead and chin) and all non-relevant areas remain dark blue, indicating strong token-to-concept alignment.

With HAAD, attention becomes more erratic and partially misaligned. While central features still attract some focus, peripheral areas like the hairline and sides of the head begin to receive attention, showing that semantic grounding is weakening.

HAAD-KV results in further disruption. HAAD-KV leads to a complete loss of structure: the attention no longer clusters around any discernible region. Instead, it is diffusely spread across the background and unrelated parts of the image, with weak, scattered activations appearing in regions such as the neck, shoulders, or background textures. The map is characterized by widespread low-intensity coloring and an absence of concentrated attention hotspots, indicating that the model has effectively lost its ability to anchor the "sks" token to any consistent visual concept.

We also compare the parameter efficiency of different methods by reporting the number of trainable parameters updated while introducing the protection into the clean images (Table 2). HAAD-KV requires the fewest updated parameters—significantly fewer than all other baselines—while still achieving top performance. This highlights its lightweight nature and confirms that perturbing only the KV parameters in the cross-attention layer is sufficient to disrupt the personalization process effectively. A more complete visualization, involving other methods considered in this paper, is provided in Supp. Material, Figure 9.

Table 2: Number of updated parameters while introducing the protection into the clean images (in millions).

AdvDM	ACE	ACE+	CAAT	HAAD	HAAD-KV
859.52	123.06	123.06	19.17	97.03	5.24

4.5 Generalization and Robustness Analysis

In this section, we study the impact of our methods on various operations: (a) purification methods (as studied in [82]) (b) prompt invariance (as studied by [53]) (c) image editing (example SDEdit[84] as studied by ACE) and (d) transferability to different models (as studied by ACE).

4.5.1 Robustness to Purification Methods

To evaluate the robustness of the protection introduced by adversarial perturbations, we follow prior work [82] to test whether common image pre-processing techniques can “purify” the perturbed images which could restore the original functionality of the diffusion model. Specifically, we apply a set of standard transformations including Gaussian noise ($\sigma = 4, 8$), Gaussian blur (kernel size 3, 5), JPEG compression (quality 20, 70), resizing (including two setups: $2\times$ up-scaling + recovering ($2\times$) and $0.5\times$ down-scaling + recovering ($0.5\times$)), and super-resolution (SR) [89] to the protected image.

We report CLIP-IQA (CI) scores under several representative purification settings in Table 3, including Gaussian noise ($\sigma = 8$), Gaussian blur (kernel size 5), JPEG ($Q = 70$), resizing with $0.5\times$ and SR. A higher CI score indicates stronger protection. HAAD-KV consistently achieves the best performance across all shown methods, while HAAD ranks second in most cases. These results suggest that both variants are robust against common purification techniques. For a more complete analysis, refer to Supp. Material, Table 10 and Figure 6.

Table 3: CLIP-IQA scores under selected purification settings.

Protection / Defense	Gaussian noise ($\sigma = 8$)	Gaussian blur (ker. 5)	JPEG ($Q = 70$)	Resizing ($0.5\times$)	SR
AdvDM	20.91	29.62	21.94	22.02	33.23
ACE	25.96	28.55	23.11	25.91	35.26
ACE+	24.68	29.35	22.89	26.06	34.76
CAAT	27.81	33.33	26.05	26.38	35.55
HAAD	28.71	33.12	28.87	26.37	35.71
HAAD-KV	29.15	33.61	29.26	27.88	36.84

4.5.2 Prompt Invariant Protection

To evaluate the impact of the robustness of our methods against different prompts, we conduct inference using a diverse set of prompts [53] that describe different contexts and poses. Here we aim to assess whether protection remains effective when the prompt is changed while keeping the protected image fixed. We perform this experiment with six different prompts: "a dslr portrait of sks person", "a photo of sks person looking at the mirror", "a photo of sks person sitting on a chair", "a photo of sks person sitting on the floor", "a photo of sks person wearing glasses", "a photo of sks person talking on the phone".

Table 4 shows the quantitative results of "a dslr portrait of sks person". Across all metrics, We can observe HAAD and HAAD-KV achieve the best or second-best performances consistently, demonstrating the strong generalization to unseen contexts. Additional results for all six inference prompts are provided in Supp. Material, Table 12 and Figure 7. These results confirm that our method generalizes effectively to diverse and unconstrained generation scenarios while remaining robust even when prompt semantics shift significantly.

Table 4: Quantitative results with a different prompt "a dslr portrait of sks person" during inference.

	"a dslr portrait of sks person"			
	CI \uparrow	CS \downarrow	FDFR \uparrow	ISM \downarrow
AdvDM	18.12	76.91	0.005	0.6366
ACE	25.96	74.47	0.010	0.5921
ACE+	25.77	75.10	0.007	0.5954
CAAT	29.48	74.88	0.060	0.5917
HAAD	29.37	73.85	0.090	0.5899
HAAD-KV	30.37	72.09	0.115	0.5682

4.5.3 Protection against Image Editing

While our primary goal was protection against concept-level customization, we also explore the potential of our method towards image editing through a preliminary study on SDEdit [84]—a popular image-to-image editing framework based on diffusion models. Although SDEdit is not explicitly designed for concept learning, it can be used to make modifications to personal images with a prompt, raising practical concerns for misuse in identity editing. Table 5 reports performance using two evaluation metrics: Multi-Scale SSIM (MS) and CLIP-SIM (CS). Our method achieves the lowest scores across both metrics, suggesting that the perturbations remain partially effective even under this distinct editing paradigm. Notably, HAAD-KV achieves the best overall results on both datasets. More qualitative comparisons are shown in Supp. Material, Figures 12, 15. These results suggest that our method is not limited to text-to-image customization, but it may also offer a generalized protection against diffusion model-based image editing. While not our main focus, this preliminary study indicates that our attack strategy retains effectiveness even under structurally guided editing, laying a foundation for future research into editing-aware protections.

4.5.4 Transferability to different models

Transferability is an essential aspect of any protection method, as real-world scenario misuse may occur across diverse diffusion model architectures and versions. If a protection is tightly coupled to a specific model, it can be easily circumvented by switching to a different backbone. To evaluate this, we assess the robustness of our method on LoRA+Dreambooth in Table 13. For this experiment, multiple versions of Stable Diffusion (i.e., v1.4, v1.5, v2.1) are used in turn to generate adversarial perturbations ("Attacker") and then evaluated on all three models ("Target"),

Table 5: Comparison of our methods with other adversarial attack methods on SDEdit. **bold** is the best, while underline indicates the second best. HAAD and HAAD-KV achieves the best performance.

	CelebA-HQ		WikiArt	
	MS ↓	CS ↓	MS ↓	CS ↓
No Attack	0.3637	79.26	0.1433	79.68
AdvDM	0.3561	77.90	0.1421	77.38
ACE	0.3393	75.53	0.1411	75.39
ACE+	0.3366	76.27	0.1352	76.47
CAAT	0.3488	77.16	0.1374	76.11
HAAD	<u>0.3360</u>	76.10	0.1293	75.85
HAAD-KV	0.3349	<u>75.58</u>	<u>0.1208</u>	<u>75.75</u>

Table 6: CLIP IQA scores for Transferability of HAAD across different versions of SD.

Target	SD1.4	SD1.5	SD2.1
No Attack	18.89	21.32	19.18
SD1.4	27.51	29.38	30.36
SD1.5	29.32	29.10	30.55
SD2.1	29.27	26.81	29.62

covering both forward and backward transfer. Across all settings, HAAD retains strong protection ability, with minimal performance drop when transferred across SD versions. This demonstrates that our attack is not tied to a specific version but generalizes well within the Stable Diffusion family. In addition, we conducted a preliminary study on Stable Diffusion 3 (SD3) [86], a recently released model based on DiT (Diffusion Transformer) architecture. Using perturbations generated on SD1.5, we apply them to SD3 and observe visual degradation in both facial identity and artistic structure (refer Supp. Material, Figure 8). While these results are not yet conclusive, they suggest that our method may have initial transferability potential beyond UNet-based backbones, laying the groundwork for further exploration. More visual results on SD 1.4–2.1 are provided in Supp. Material, Table 14.

5 Conclusion

In this work, by exploiting the semantic structure of the latent space of the U-Net (h -space), we introduced HAAD and its more efficient variant HAAD-KV, two simple yet effective adversarial strategies designed to protect against unauthorized few-shot image personalization by diffusion models. HAAD, by focusing on the whole h -space, disrupts the alignment between user-specific tokens and visual concepts. HAAD-KV, the computationally efficient variant, finds the perturbation based only on the KV parameters of the cross-attention layer in h -space, achieving stronger protection with fewer trainable parameters. Extensive experiments across diverse personalization frameworks (LoRA+DB, CD) and datasets (CelebA-HQ, WikiArt) showed that our approaches consistently outperform existing methods, both in semantic distortion and perceptual degradation. We also conducted comprehensive robustness studies under standard image purification transformations, varying prompts, and multiple Stable Diffusion versions (including image editing as an additional use-case), validating the generalizability and transferability of our methods. Preliminary results on SD3 and SDEdit suggested that our approach may extend to broader generative pipelines, opening new directions for editing-aware and architecture-agnostic defenses. Overall, HAAD and HAAD-KV demonstrated that leveraging latent semantics offers a promising and efficient pathway towards safeguarding personal content in generative models.

Acknowledgements

Xide Xu acknowledges the Chinese Scholarship Council (CSC) grant No.202306310064. This work is supported by Grant PID2022-143257NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A Way of Making Europa", the Departament de Recerca i Universitats from Generalitat de Catalunya with reference 2021SGR01499, and the Generalitat de Catalunya CERCA Program.

References

- [1] George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR)*, 2014 14th International Conference on, pages 417–422. IEEE, 2014.

- [2] George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR)*, 2014 6th International Conference of, pages 312–318. IEEE, 2014.
- [3] Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.
- [4] FirstName Alpher. Frob-nication. *IEEE TPAMI*, 12(1):234–778, 2002.
- [5] FirstName Alpher and FirstName Gamow. Can a computer frobnicate? In *CVPR*, pages 234–778, 2005.
- [6] Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors. *Computer Vision – ECCV 2022*. Springer, 2022.
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020.
- [8] Stefano Ermon Jiaming Song, Chenlin Meng. Denoising diffusion implicit models. In *Proc. of ICLR*, 2021.
- [9] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. of ICLR*, 2021.
- [10] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proc. of NeurIPS*, 2021.
- [11] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Proc. of NeurIPS*, 2021.
- [12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. of NeurIPS*, 2022.
- [13] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of CVPR*, pages 10684–10695, 2022.
- [15] StabilityAI. Stable diffusion. <https://huggingface.co/stabilityai>, 2024. Hugging Face Model Hub.
- [16] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. on Graphics*, 42(4):1–10, 2023.
- [17] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, , and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. In *Proc. of ACL*, volume 1, pages 5644—5659, 2023.
- [18] Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, Tivatis Sim, and Kenji Kawaguchi. Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models. In *Proc. of ECCV*, 2024.
- [19] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2023.
- [20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *Proc. of ICLR*, 2023.
- [21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proc. of CVPR*, pages 6038–6047, 2023.
- [22] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Proc. of SIGGRAPH*, 2022.
- [23] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proc. of ICLR*, 2023.
- [24] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proc. of CVPR*, pages 20908–20918, 2023.
- [25] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proc. of ICCV*, pages 22819–22829, 2023.
- [26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

- [27] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127, 2023.
- [28] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. arXiv preprint arXiv:2312.06662, 2023.
- [29] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. Diffusion models for medical anomaly detection. In Proc. of MICCAI, pages 35–45, 2022.
- [30] Momina Masood, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, and Aun Irtaza. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Applied Intelligence, 53:3974–4026, 2023.
- [31] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. In Proc. of USENIX Conference on Security Symposium, pages 2187–2204, 2023.
- [32] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In 2024 IEEE Symposium on Security and Privacy (SP). IEEE, 2024.
- [33] Robert Chesney and Danielle Citron. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Affairs, 98(1):147–155, 2019.
- [34] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In Proc. of ICLR, 2023.
- [35] Chen Henry Wu and Fernando de la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In Proc. of ICCV, pages 7378–7387, 2023.
- [36] Rene Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Grasshof, Sami S. Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In Proc. of Automatic Face and Gesture Recognition (FG), pages 1–9, 2024.
- [37] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In Proc. of CVPR, pages 12006–12016, 2024.
- [38] Ayodeji Ijishakin, Ming Liang Ang, Levente Baljer, Daniel Chee Hian Tan, Hugo Laurence Fry, Ahmed Abdulaal, Aengus Lynch, and James H. Cole. H-space sparse autoencoders. In NeurIPS, Workshop on Safe Generative AI, 2024.
- [39] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. In Proc. of CVPR, pages 24209–24218, 2024.
- [40] Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In Proc. of ECCV, 2024.
- [41] Chenyu Zhang, Mingwang Hu, Wenhui Li, and Lanjun Wang. Adversarial attacks and defenses on text-to-image diffusion models: A survey. Information Fusion, 114(102701):1–15, 2025.
- [42] Xide Xu, Muhammad Atif Butt, Sandesh Kamath, and Bogdan Raducanu. Privacy protection in personalized diffusion models via targeted cross-attention adversarial attack. arXiv preprint arXiv:2411.16437, 2024.
- [43] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In Proc. of ICLR, 2019.
- [44] Shuai Jia, Bangjie Yin, Taiping Yao, Shouhong Ding, Chunhua Shen, Xiaokang Yang, and Chao Ma. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. In Proc. of NeurIPS, 2022.
- [45] Jing Yang, Runping Xi, Yingxin Lai, Xun Lin, and Zitong Yu. Ddap: Dual-domain anti-personalization against text-to-image diffusion models. In Proc. of IEEE Int’l. Joint Conference on Biometrics (IJCB), pages 1–10, 2024.
- [46] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In Proc. of CVPR, pages 1932–1941, 2023.
- [47] 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 Proc. of CVPR, pages 22500–22510, 2023.
- [48] 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. In Proc. of ICLR, 2023.

- [49] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proc. of ICCV*, pages 7323–7334, 2023.
- [50] 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. In *Proc. of ICLR*, 2024.
- [51] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proc. of CVPR*, pages 8543–8552, 2024.
- [52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yeal Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proc. of CVPR*, pages 6527–6536, 2024.
- [53] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proc. of ICCV*, pages 2116–2127, 2023.
- [54] Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proc. of CVPR*, pages 24219–24228, 2024.
- [55] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. of ICLR*, 2015.
- [56] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. of ICLR*, 2018.
- [57] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proc. of ICML*, volume 119, pages 2206–2216, 2020.
- [58] Yunrui Yu, Xitong Gao, and Cheng-Zhong Xu. Lfeat: Piercing through adversarial defenses with latent features. In *Proc. of CVPR*, pages 5735–5745, 2021.
- [59] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.
- [60] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *Proc. of ICML*, volume 202, pages 20763–20786, 2023.
- [61] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683v1*, 2023.
- [62] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *Proc. of ICML*, volume 202, pages 29894–29918, 2023.
- [63] Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *Proc. of CVPR*, pages 24534–24543, 2024.
- [64] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. In *Proc. of NeurIPS*, 2023.
- [65] Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *Proc. of ICLR*, 2024.
- [66] Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proc. of CVPR*, pages 12047–12056, 2024.
- [67] Yisu Liu, Jinyang An, Wanqian Zhang, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. *arXiv preprint arXiv:2405.20584*, 2024.
- [68] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. of ICCV*, pages 3730–3738, 2015.
- [69] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proc. of ICLR*, 2018.
- [70] 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 (CVPR)*, pages 10684–10695, June 2022.

- [71] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015.
- [72] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, volume 2, pages 1398–1402. Ieee, 2003.
- [73] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 2555–2563, 2023.
- [74] 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. CVPR, 2023.
- [75] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multilevel face localisation in the wild. In Proc. of CVPR, pages 5203–5212, 2020.
- [76] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proc. of CVPR, pages 4690–4699, 2019.
- [77] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Proc. of NeurIPS, 2017.
- [78] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: unsupervised estimation of face image quality based on stochastic embedding robustness. In Proc. of CVPR, pages 5650–5659, 2020.
- [79] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.
- [80] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 10609–10619. IEEE, 2022.
- [81] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In TProc. of ICLR, 2022.
- [82] Boyang Zheng, Chumeng Liang, and Xiaoyu Wu. Targeted attack improves protection against unauthorized diffusion customization. In Proc. of ICLR, 2025.
- [83] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855, 2015.
- [84] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In Proc. of ICLR, 2022.
- [85] Zhengyue Zhao, Jinhao Duan, Xing Hu, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation, 2024.
- [86] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first International Conference on Machine Learning, 2024.
- [87] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative AI. In Proc. of ICLR, 2025.
- [88] Jinyang An, Wanqian Zhang, Dayan Wu, Zheng Lin, Jingzi Gu, and Weiping Wang. Sd4privacy: Exploiting stable diffusion for protecting facial privacy. In IEEE International Conference on Multimedia and Expo, ICME 2024, Niagara Falls, ON, Canada, July 15-19, 2024, pages 1–6. IEEE, 2024.
- [89] Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. IEEE Transactions on Image Processing, 29:1711–1724, 2019.
- [90] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative ai. arXiv preprint arXiv:2406.12027, 2024.
- [91] Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. IMPRESS: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative AI. In Thirty-seventh Conference on Neural Information Processing Systems, 2023.
- [92] Louis Leon Thurstone. A law of comparative judgment. Psychological Review, 34:273–286, 1927.

6 Supplementary Material

The supplementary contains experimental details and additional results for the experiments in the main paper.

6.1 Experimental Details : Hyperparameters

Table 7: Hyperparameters for different attack methods. The parameters for all the methods are set to their default settings. Noise budget of $\eta = 4/255$ was used in all our study.

Parameters	AdvDM	CAAT	ACE	HAAD
train steps	1000	250	50	250
learning rate	1×10^{-4}	1×10^{-5}	5×10^{-6}	1×10^{-5}
α	2/255	5×10^{-3}	5×10^{-3}	5×10^{-3}

Table 8: Hyperparameters for different custom diffusion models, with their default settings.

Parameters	LoRA+DB	CD
train steps	1000	250
learning rate	5×10^{-5}	1×10^{-5}
batchsize	1	2
LoRA rank	4	-

6.2 Results with different noise budgets η (HAAD vs HAAD-KV)

An increased noise budget (η) could result in perturbations that become more perceptible to the human eye. Figure 5 and Table 9 show the effectiveness of attack with different noise budgets. At a low noise budget of 4/255 we find it hard to observe any visible change in the perturbation added to the image, hence, considered imperceptible to humans, but still creates noticeable changes to features in the generated image. At noise budget of greater than 8/255, the perturbations start becoming more pronounced (observable with zoom in Figure 5) and resulting in the creation of faint artificial patterns (e.g., grid-like stripes) in the generated images. At a higher noise budget of 16/255, the adversarial samples exhibit the largest change, resulting in the generated images to be almost completely unrecognizable. This noise level ensures the highest level of privacy protection, as the images become highly blurred and indistinct.

The results show that HAAD-KV offers better protection than HAAD. At the same time, even with a small noise budget of 4/255, our methods remains highly effective and are comparable to other methods operating at higher noise levels ($> 8/255$).

Table 9: A quantitative comparison of HAAD and HAAD-KV on LoRA+DB with CelebA-HQ.

	CI \uparrow		FDFR \uparrow		ISM \downarrow	
	HAAD	HAAD-KV	HAAD	HAAD-KV	HAAD	HAAD-KV
4/255	29.10	29.52	0.085	0.100	0.5175	0.5083
8/255	37.70	40.01	0.110	0.180	0.4517	0.4384
16/255	46.11	48.41	0.690	0.820	0.2905	0.2357

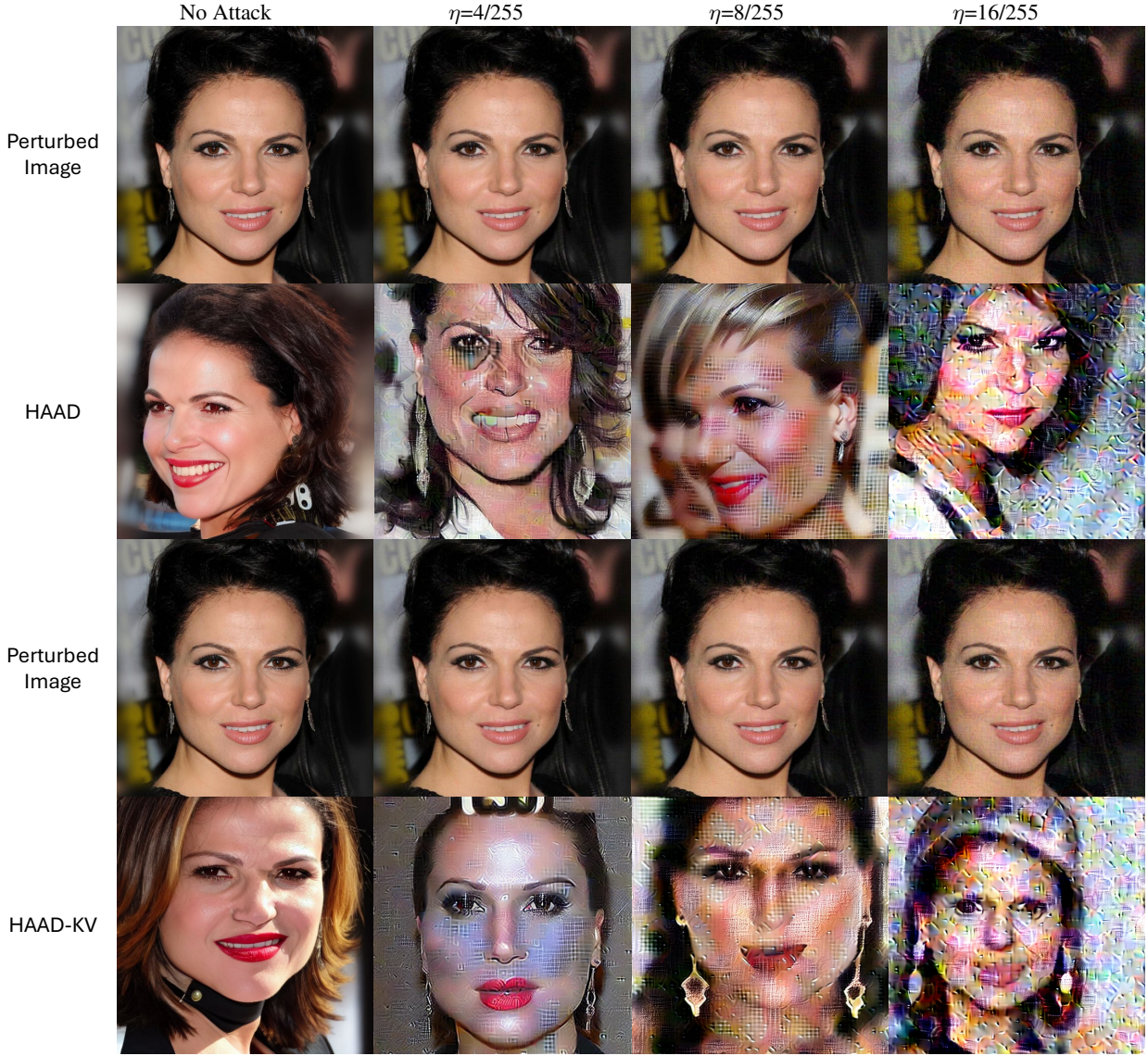


Figure 5: Comparison between HAAD (first two rows) and HAAD-KV (second two rows) for varying noise budgets of $\{0, 4/255, 8/255, 16/255\}$ (from left to right). We observe a consistent degradation of the generated images (using LoRA+DB as customization method).

6.3 Qualitative and quantitative results comparison of robustness between adversarial methods.

In this section, we show the effect of purification methods on the protection added by different adversarial methods. We include the various tests applied in [63, 82]. These results corroborate our conclusion: HAAD-KV not only introduces visually imperceptible perturbations but also withstands “purification” techniques that could otherwise neutralize fixed-pattern attacks like ACE or CAAT. Figure 6 provides a visual comparison of the outputs from different methods after applying the corresponding purification method. It can be seen that baseline methods often are able to recover partially recognizable content, while HAAD-KV protection consistently leads to distorted, incoherent generations — demonstrating its robustness and practical effectiveness.

Table 10: Quantitative results for different attack methods under different purification.

CLIP IQA (CI) \uparrow									
Defense	Gaussian noise		Gaussian blur		JPEG		Resizing		SR
Parameter	$\sigma = 4$	$\sigma = 8$	$K = 3$	$K = 5$	$Q = 20$	$Q = 70$	$2\times$	$0.5\times$	
AdvDM	22.71	20.91	24.88	29.62	32.91	21.94	26.23	22.02	33.23
ACE	26.16	25.96	26.12	28.55	33.65	23.11	27.54	25.91	35.26
ACE+	23.18	24.68	25.78	29.35	33.34	22.89	28.78	26.06	34.76
CAAT	<u>28.91</u>	27.81	31.90	<u>33.33</u>	34.84	26.05	32.00	<u>26.38</u>	35.55
HAAD	28.86	<u>28.71</u>	30.37	33.12	<u>36.18</u>	<u>28.87</u>	<u>33.92</u>	26.37	<u>35.71</u>
HAAD-KV	31.51	29.15	<u>31.16</u>	33.61	39.62	29.26	34.85	27.88	36.84
CLIP SIM (CS) \downarrow									
AdvDM	82.41	81.51	80.24	78.83	78.94	80.43	79.18	81.31	77.72
ACE	80.53	80.24	78.56	77.67	78.69	81.05	80.49	80.85	77.54
ACE+	81.07	80.79	78.94	77.81	78.87	80.58	79.78	80.77	77.94
CAAT	75.41	75.68	<u>75.57</u>	74.67	78.28	79.96	78.27	79.39	76.14
HAAD	<u>74.63</u>	<u>74.19</u>	75.91	<u>74.58</u>	74.57	<u>78.89</u>	<u>78.12</u>	<u>77.03</u>	<u>75.31</u>
HAAD-KV	68.95	72.53	75.41	74.13	<u>75.11</u>	77.91	76.93	68.98	72.63

We even evaluated the robustness of our protection against recent purification techniques proposed in [90], specifically Noisy Upscaling [90] and Impress [91]. The goal is to determine if these purification methods could remove or reduce the protection by the adversarial perturbation, thereby restoring the image’s utility for personalization. Table 11, summarizes the impact of these methods on the image quality. We observe that, Noisy Scaling severely degrades the image quality, which is evidenced by the sharp drop in SSIM (0.3463) and PSNR (23.36). In contrast, Impress better preserves the perceptual quality, though it reduces fidelity compared to the original protected image. Notably, we used the purified images from both methods as inputs for LoRA+DreamBooth. In both cases, the personalization process failed to faithfully reconstruct the target identity. This demonstrates that HAAD-KV perturbations are robust, withstanding purification attempts and effectively preventing unauthorized customization.

Table 11: Evaluation of purification techniques against HAAD-KV protection. While purification methods like Noisy Upscaling and Impress alter image quality metrics, HAAD-KV remains effective, preventing personalization, even after purification efforts.

Method	SSIM \uparrow	PSNR \uparrow	CLIP-SIM(CS) \downarrow
No Attack (Original)	1.0000	∞	83.13
HAAD-KV	0.9862	59.95	71.91
<i>Purification applied to HAAD-KV protected Image:</i>			
Noisy Upscaling	0.3463	23.36	77.62
Impress	0.9209	32.84	74.64

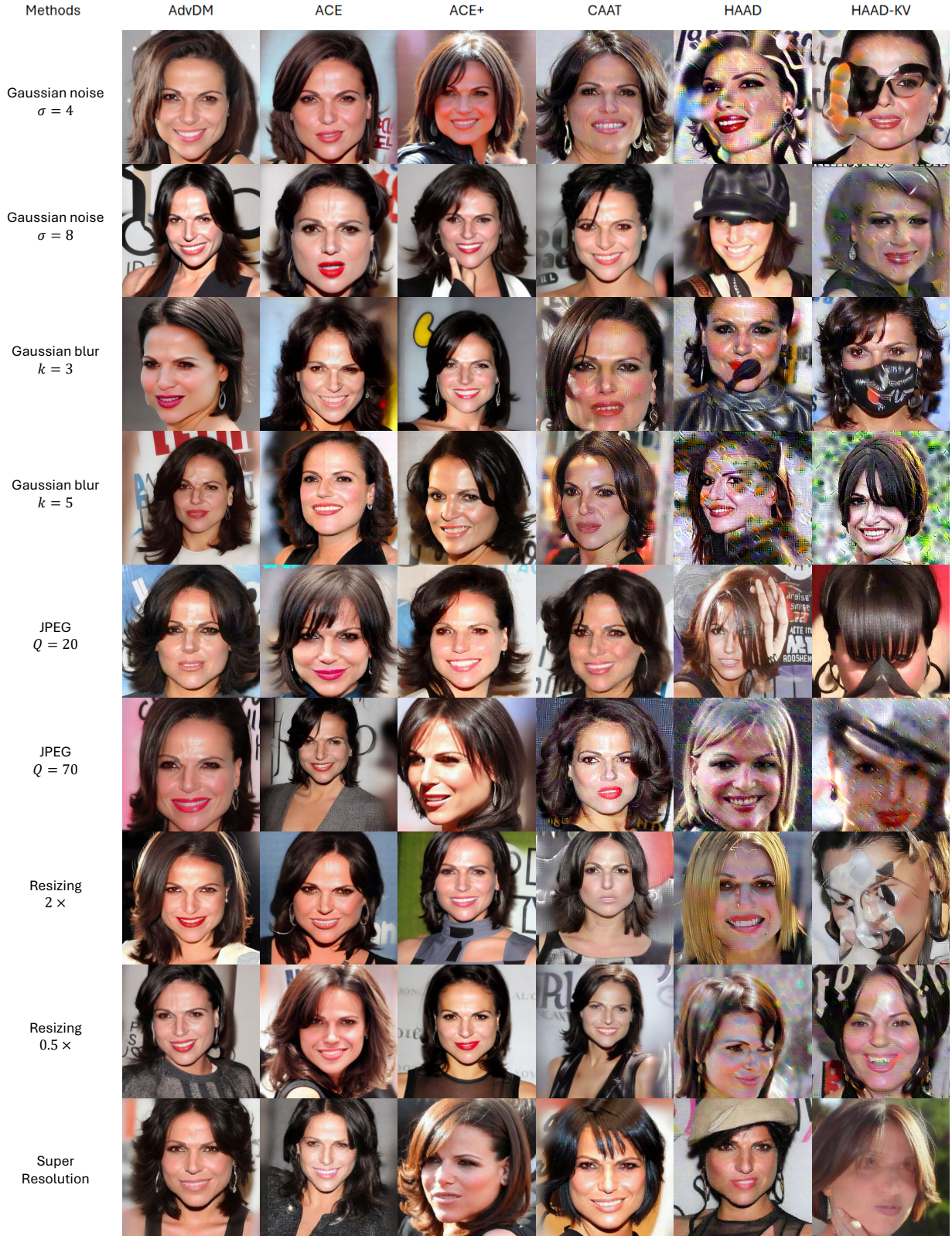


Figure 6: Qualitative results for different attack methods under different purification.

6.4 Qualitative and quantitative results with different prompts.

Table 12: Prompt-Invariant Performance across Six Prompts. Quantitative results (CI, CS, FDFR, ISM) for all tested prompts, evaluating the robustness of defense methods under prompt variation. **bold** is the best, while underline indicates the second best. HAAD and HAAD-KV achieves the best performance.

	"a dslr portrait of sks person"				"a photo of sks person looking at the mirror"			
	CI \uparrow	CS \downarrow	FDFR \uparrow	ISM \downarrow	CI \uparrow	CS \downarrow	FDFR \uparrow	ISM \downarrow
AdvDM	18.12	76.91	0.005	0.6366	23.43	74.75	0.050	0.5568
ACE	25.96	74.47	0.010	0.5921	29.60	76.09	0.025	0.5923
ACE+	25.77	75.10	0.007	0.5954	29.25	75.80	0.010	0.5908
CAAT	<u>29.48</u>	74.88	0.060	0.5917	29.04	74.40	0.060	0.5388
HAAD	29.37	<u>73.85</u>	0.090	<u>0.5899</u>	31.51	69.76	<u>0.115</u>	0.5118
HAAD-KV	30.37	72.09	0.115	0.5682	<u>29.76</u>	<u>72.65</u>	0.145	<u>0.5304</u>
	"a photo of sks person sitting on a chair"				"a photo of sks person sitting on the floor"			
	CI \uparrow	CS \downarrow	FDFR \uparrow	ISM \downarrow	CI \uparrow	CS \downarrow	FDFR \uparrow	ISM \downarrow
AdvDM	28.12	70.97	0.150	0.4338	33.24	66.76	0.105	0.5168
ACE	33.73	66.96	0.255	0.3944	34.81	63.67	0.215	0.4728
ACE+	32.57	67.52	0.220	0.4086	34.02	63.89	0.190	0.4871
CAAT	33.96	66.73	0.280	0.3853	35.15	62.43	0.260	0.4440
HAAD	<u>35.36</u>	<u>64.98</u>	<u>0.312</u>	<u>0.3712</u>	<u>35.56</u>	<u>61.02</u>	<u>0.305</u>	<u>0.4117</u>
HAAD-KV	35.68	64.13	0.345	0.3508	35.82	60.49	0.320	0.4051
	"a photo of sks person wearing glasses"				"a photo of sks person talking on the phone"			
	CI \uparrow	CS \downarrow	FDFR \uparrow	ISM \downarrow	CI \uparrow	CS \downarrow	FDFR \uparrow	ISM \downarrow
AdvDM	22.87	75.78	0.010	0.6076	20.78	78.67	0.025	0.7341
ACE	23.55	74.21	0.050	0.5745	31.58	73.69	0.095	0.6504
ACE+	23.21	75.02	0.035	0.5924	30.24	74.13	0.070	0.6697
CAAT	23.56	73.75	0.105	0.5554	32.81	73.43	0.115	0.6332
HAAD	<u>23.99</u>	<u>73.06</u>	<u>0.155</u>	<u>0.5302</u>	<u>33.43</u>	<u>72.56</u>	<u>0.190</u>	<u>0.6279</u>
HAAD-KV	24.18	72.87	0.185	0.5217	33.51	71.78	0.235	0.6261

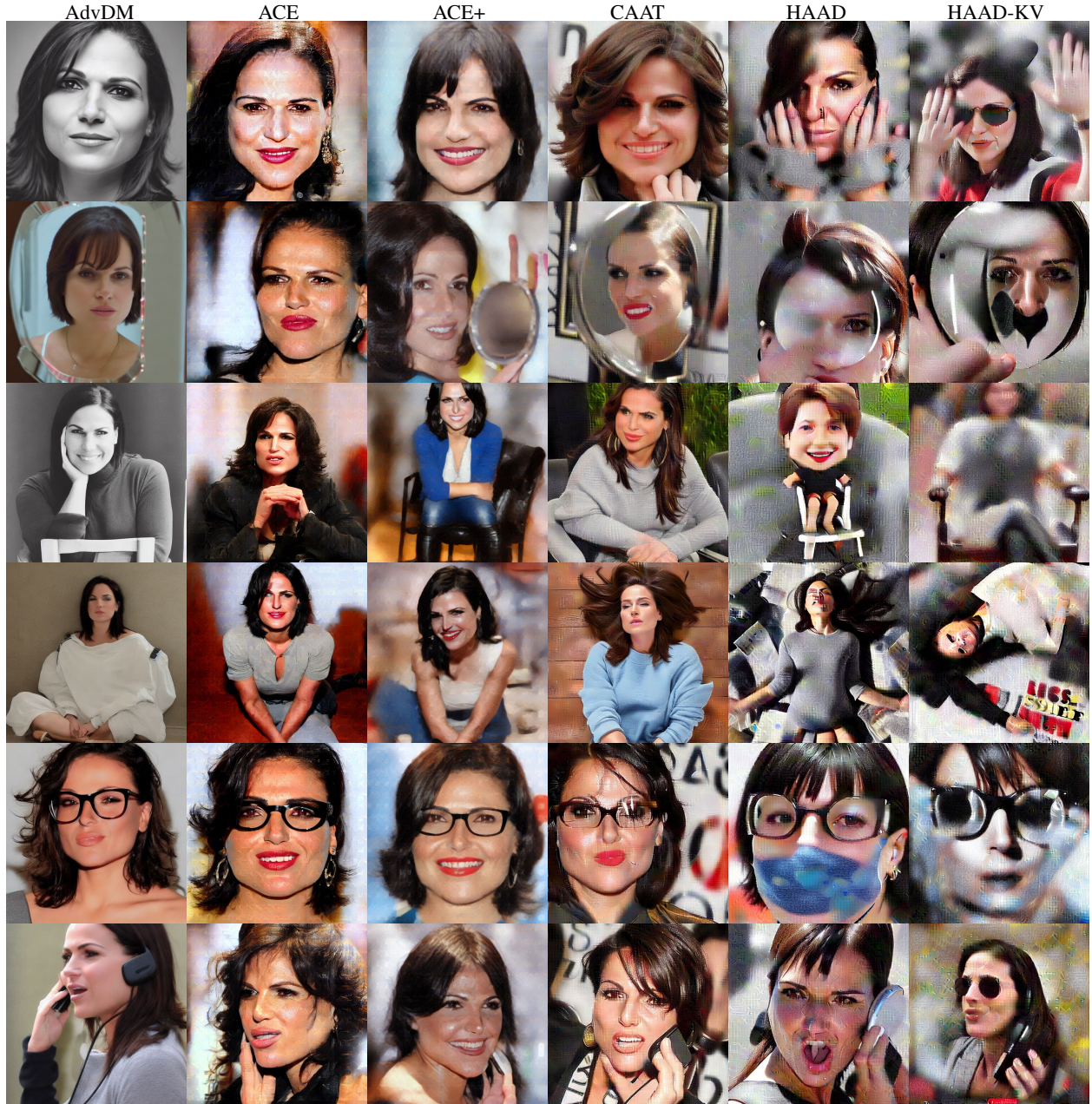


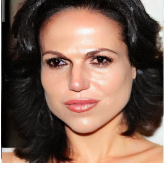
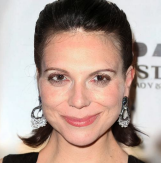
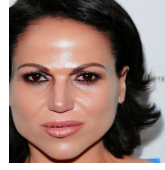
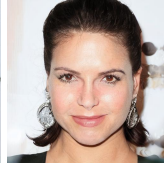

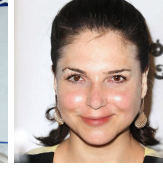
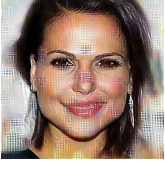
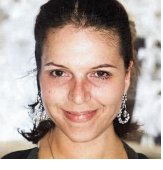
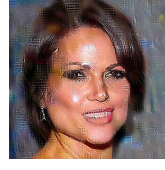

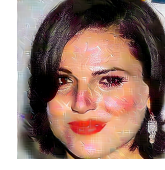
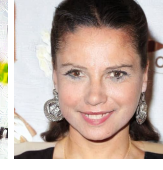
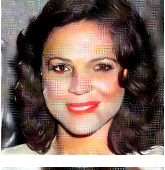

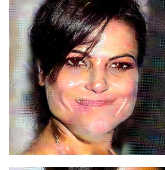
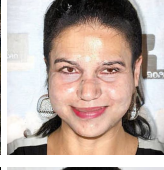
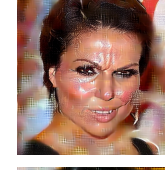
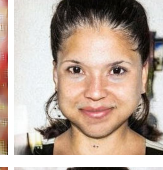
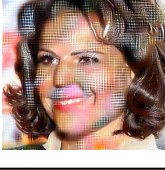
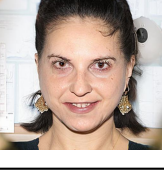
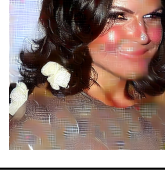
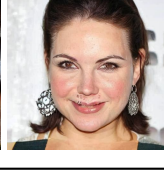
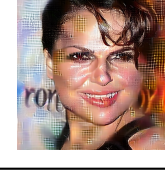
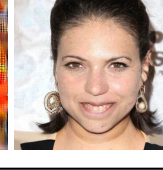
Figure 7: Results with different prompts during inference. (Row 1) "a dslr portrait of sks person", (Row 2) "a photo of sks person looking at the mirror", (Row 3) "a photo of sks person sitting on a chair", (Row 4) "a photo of sks person sitting on the floor", (Row 5) "a photo of sks person wearing glasses", (Row 6) "a photo of sks person talking on the phone".

6.5 Transferability study using HAAD : Qualitative results.

Table 13: Transferability of HAAD across different versions of SD.

Target	SD1.4			SD1.5			SD2.1		
Attacker	LoRA+DB	SDEdit		LoRA+DB	SDEdit		LoRA+DB	SDEdit	
	CI \uparrow	MS \downarrow	CS \downarrow	CI \uparrow	MS \downarrow	CS \downarrow	CI \uparrow	MS \downarrow	CS \downarrow
No Attack	18.89	0.3694	75.01	21.32	0.3637	79.26	19.18	0.3782	74.67
SD1.4	27.51	0.3498	73.56	29.38	0.3444	73.62	30.36	0.3587	72.93
SD1.5	29.32	0.3479	73.23	29.10	0.3360	75.80	30.55	0.3568	72.85
SD2.1	29.27	0.3455	73.18	26.81	0.3435	73.23	29.62	0.3561	72.72

Table 14: Transferability of HAAD among the different versions of SD. We observe consistent degradation for all the methods.

Target	SD1.4		SD1.5		SD2.1	
Attacker	LoRA+DB	SDEdit	LoRA+DB	SDEdit	LoRA+DB	SDEdit
No Attack						
SD1.4						
SD1.5						
SD2.1						

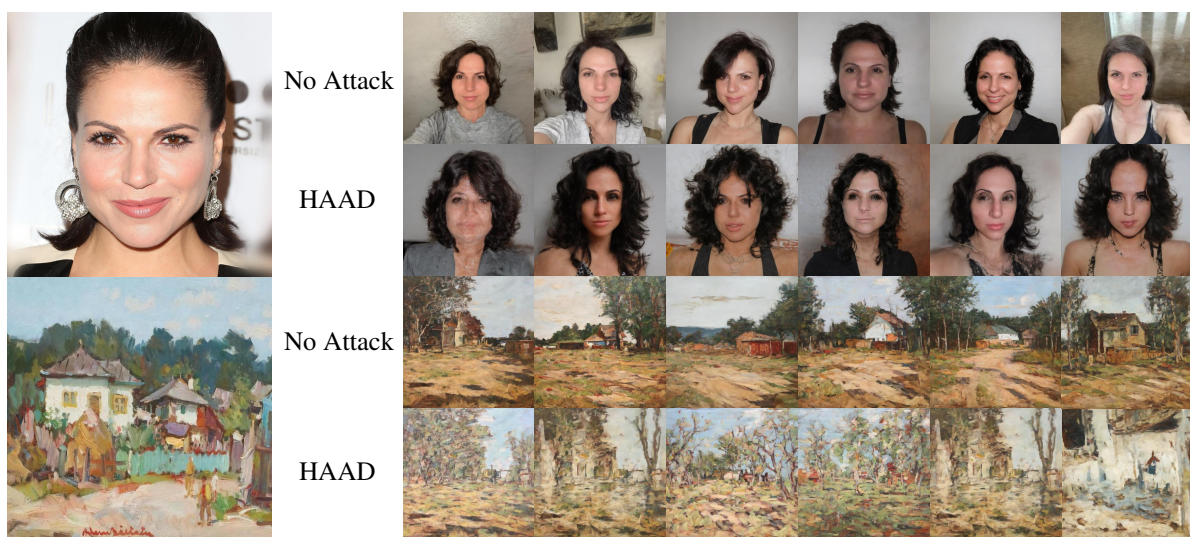


Figure 8: SD1.5 is used as the ‘Attacker’ and SD3 as the ‘Target’ model (used to generate the shown images). The figure highlights perceptible distortions in facial features, especially the eyes, while art images undergo distinct stylistic alterations.

6.6 Visualization of cross-attention maps for all the attack methods.

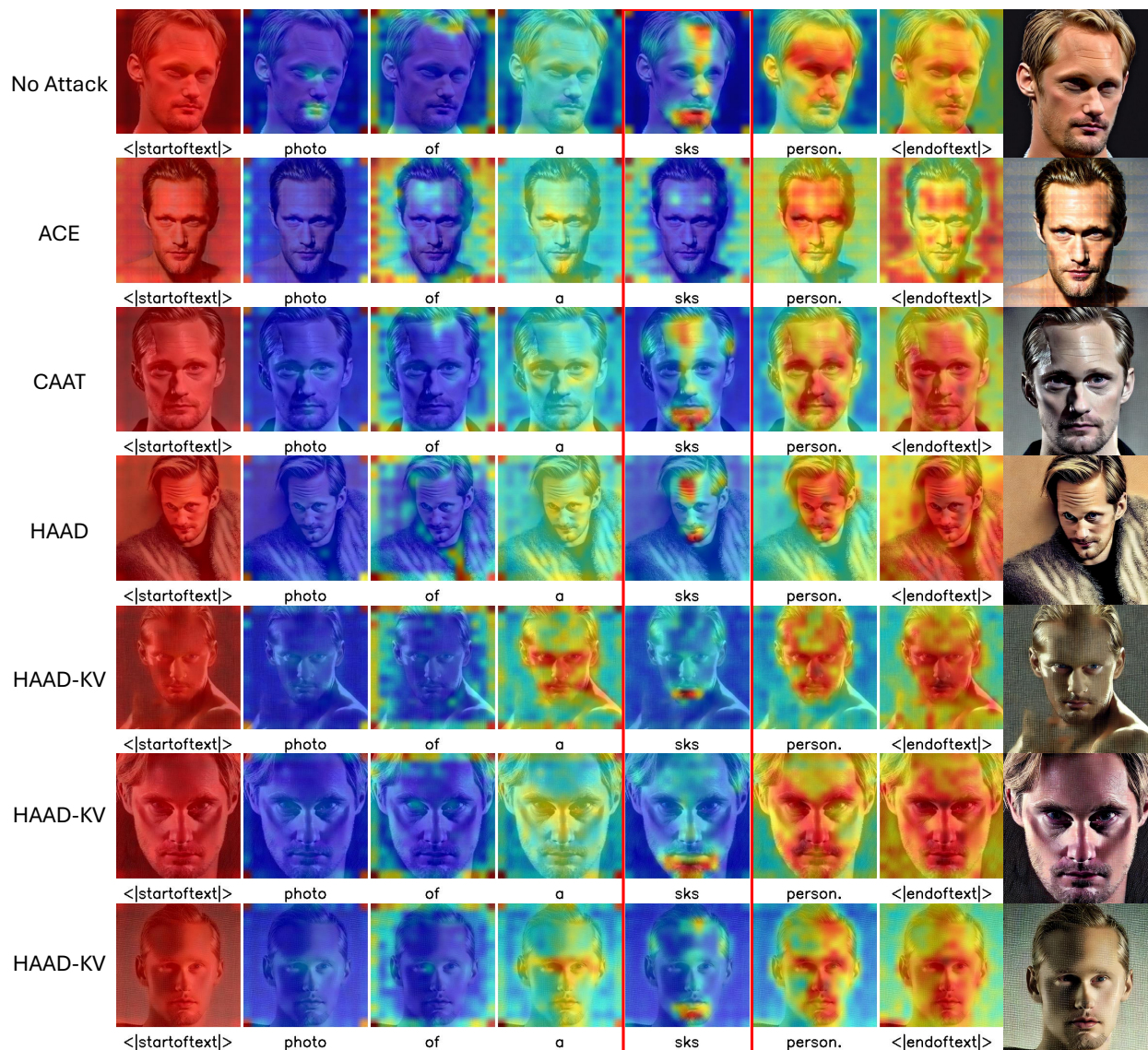


Figure 9: Visualization of the cross attention map of each token during inference time for different adversarial attack methods on LoRA+DB with a sample of CelebA-HQ dataset.

6.7 Theoretical Framework.

***h*-space Perturbation and Semantic Misalignment.** Let W_h denote the features (weights) of the *h*-space of the U-Net, which encodes high-level semantic content during the diffusion process. For the HAAD attack, we iteratively compute adversarial perturbation for the protection of the input image x with projected gradient descent (PGD) which maximizes the reconstruction loss \mathcal{L}_{LDM} . Each step updates the semantic features as given by:

$$W'_h = W_h + l \nabla_{W_h} \mathcal{L}_{\text{LDM}}(x), \quad (3)$$

where, l is the learning rate. The final protected image is denoted as $x' = x + \delta$, and its corresponding semantic features are encoded in W'_h .

This update induces a linear shift in the *h*-space, modifying the intermediate representation in a direction aligned with ∇_{W_h} , which was adversarially chosen to degrade the semantic consistency. Since the cross-attention mechanism depends on the alignment between the textual prompt and the latent features in the *h*-space, this shift weakens the model’s ability to associate prompt tokens (e.g., “sks”) with relevant visual features, which finally affects the image generation process.

When an attacker trains a new personalization model with the protected image x' , it starts to operate on the clean semantic representation encoded in W_h . Afterwards, at the end of the personalization process, the resulting *h*-space parameters are denoted as W''_h . Because the training began with x' and W_h , we expect: $W''_h \approx W'_h$, implying that the learned representation converges towards a similar semantically misaligned features (weights) (W'_h). This is straightforward since x' carries the semantic misalignment information encoded in the perturbation.

Consequently, at inference time, the model’s attention fails to bind the user-specific token (e.g., “sks”) to the correct semantic features, due to the distortion introduced by the perturbation $\Delta W_h = W''_h - W_h$. This semantic misalignment in the attention mechanism results in visibly degraded or incoherent image generations. Thus, perturbations in *h*-space not only alter intermediate representations but also disrupt the prompt-to-concept mapping, offering an effective and semantically grounded defense against unauthorized personalization.

The role of KV layers towards Perturbation and Semantic Misalignment. HAAD-KV extends the above intuition by focusing the adversarial update specifically on the *cross-attention layers* inside the *h*-space block, and only modifying the key W_K and value W_V projection matrices.

During few-shot personalization, these matrices adapt to encode how each visual element should respond to the tokenized concept. HAAD-KV injects perturbations only into W_K and W_V , thereby corrupting the model’s ability to compute meaningful attention maps. In other words, the perturbed matrices W'_K and W'_V are given by:

$$\begin{aligned} W'_K &= W_K + l \nabla_{W_K} \mathcal{L}_{\text{LDM}}(x), \\ W'_V &= W_V + l \nabla_{W_V} \mathcal{L}_{\text{LDM}}(x) \end{aligned}$$

This formulation is similar to equation (3). As a result, the perturbed attention becomes:

$$\text{Attention}(Q, K', V') = \text{softmax} \left(\frac{QK'^T}{\sqrt{d}} \right) V'$$

where, the K' and V' are the resultant perturbed matrices, induced by W'_K and W'_V , respectively. Since this misalignment occurs at the point where identity tokens are explicitly bound to visual representations, the model fails to personalize effectively during training.

In conclusion, HAAD-KV thus offers a focused and computationally efficient attack: by perturbing only a small fraction ($\sim 5\%$ of the total parameters in *h*-space) of model parameters (those with high semantic content), it results in maximum distortion of the personalization mechanism with minimal perceptual footprint.

Analysis of representational disruption in *h*-space. To quantitatively evaluate that our method alters the core semantic structure, we perform a Principal Component Analysis (PCA) on the activations within the *h*-space. This analysis aims to measure the alignment of the primary directions of variance between representations of clean and protected images. We computed the cosine similarity between the top- k principal components derived from clean image activations and those from images protected by HAAD-KV.

Description of the PCA method: PCA identifies the orthogonal directions, or Principal Components (PCs), that capture the maximum variance within the *h*-space activations. The “top- k components” refer to the first k PCs that represent the most significant semantic patterns in the data. We then use cosine similarity to measure the alignment between the PCs

derived from clean images with those from their protected counterparts. A cosine similarity value close to 1.0 implies that the components are highly aligned, meaning the clean and protected images share the same core semantic structure. Conversely, a value close to 0.0 indicates orthogonality, meaning their semantic structures are fundamentally different and uncorrelated.

The results presented in Table 15, clearly reveal a systematic and noise budget dependent impact of HAAD-KV. At a low perturbation budget of $\eta = 4/255$, we already observe a significant drop in alignment, with the cosine similarity for the top-5 components being 0.2961. This indicates that even at low perturbation budget the protection begins to alter the primary semantic features. As the perturbation budget increases to $\eta = 8/255$ and $\eta = 16/255$, respectively, this trend is further amplified. For instance, at $\eta = 16/255$, the similarity of the top-5 components drops to 0.1382. This demonstrates that larger protection budgets lead to a greater divergence in semantic representation. Furthermore, this effect is consistent when more components were considered (i.e., k is increased from 5 to 50), confirming that the semantic shift is not limited to only the top few dominant features but across the semantic latent space.

Table 15: Cosine similarity between the top- k principal components of h-space activations for clean versus HAAD-KV protected images. The consistently low similarity scores across different perturbation budgets (η) demonstrate a significant semantic drift induced by HAAD-KV.

Perturbation (η)	Top- k Principal Components			
	$k = 5$	$k = 10$	$k = 20$	$k = 50$
4/255	0.2961	0.1413	0.0668	0.0243
8/255	0.2171	0.1111	0.0603	0.0245
16/255	0.1382	0.0672	0.0351	0.0135

6.8 User study.

To empirically verify that the injected noise is imperceptible to humans, we carried out a user study with the CelebA-HQ dataset, containing face images. For each noise budget $\eta \in \{4/255, 8/255, 16/255\}$ we sampled **10** identities, and for every identity selected **3** high-resolution photographs, yielding **30** original-perturbed pairs per η value. Twenty-six volunteers ($26 \times 30 = 780$ judgements per budget) were shown each pair side-by-side and asked “Which image is perturbed?” with three choices: (A) first, (B) second, or (C) “both images are the same.”

Table 16 reports the scores obtained for the three noise budgets, including the **Error Rate** and **Z-score** derived from Thurstone’s Case V model[92]. In this study, the Error Rate represents the fraction of times volunteers failed to correctly identify the perturbed image (Accuracy(%) is given by $\{100 - (\text{Error Rate} \times 100)\}\%$). A higher error rate signifies that the introduced noise is less perceptible(imperceptible) and is hard to be detected by human eyes. The Z-score quantifies perceptual discriminability, with a higher value indicating easier detection. The **Standard Deviation (STD)** of the Z-score measures the level of agreement among the participants, where a lower STD indicates a more consistent perceptual experience across the group.

The analysis reveals a strong correspondence between the metrics. Specifically, for $\eta = 16/255$, a high positive Z-score of “1.23” indicates that the perturbation was clearly perceptible, a conclusion strongly supported by the very low error rate of “0.117” (88.3% accuracy) and low STD (0.29). And the low STD (0.29) for this budget shows a strong consensus that many were able to detect the perturbation in the image. For $\eta = 8/255$, the Z-score is near zero (“0.28”) suggesting discrimination was at a threshold level with a mix of both partial detection and low perception, which is corroborated by the high STD (“0.34”) and a moderate error rate of “0.392” (60.8% accuracy). Critically, for $\eta = 4/255$, the negative Z-score of “-0.39” confirms the images were perceptually indistinguishable. This is strongly demonstrated by the error rate of “0.650”, which translates to an accuracy of 35%—statistically identical to random chance for this 3-choice task. The low STD (“0.25”) for this budget further shows a strong consensus that no difference could be spotted. This confirms that at the strictest budget (4/255), participants performed at chance level, indicating that the perturbations are visually indistinguishable in practice.

Noise budget η	Mean Z-Score	STD	Error Rate
4/255	-0.39	0.25	0.650
8/255	0.28	0.34	0.392
16/255	1.23	0.29	0.117

Table 16: Thurstone Case V results and corresponding error rates. The Z-scores and error rates provide complementary evidence for the (in)discriminability of perturbations.

6.9 Qualitative results for different attack methods under different customization models.

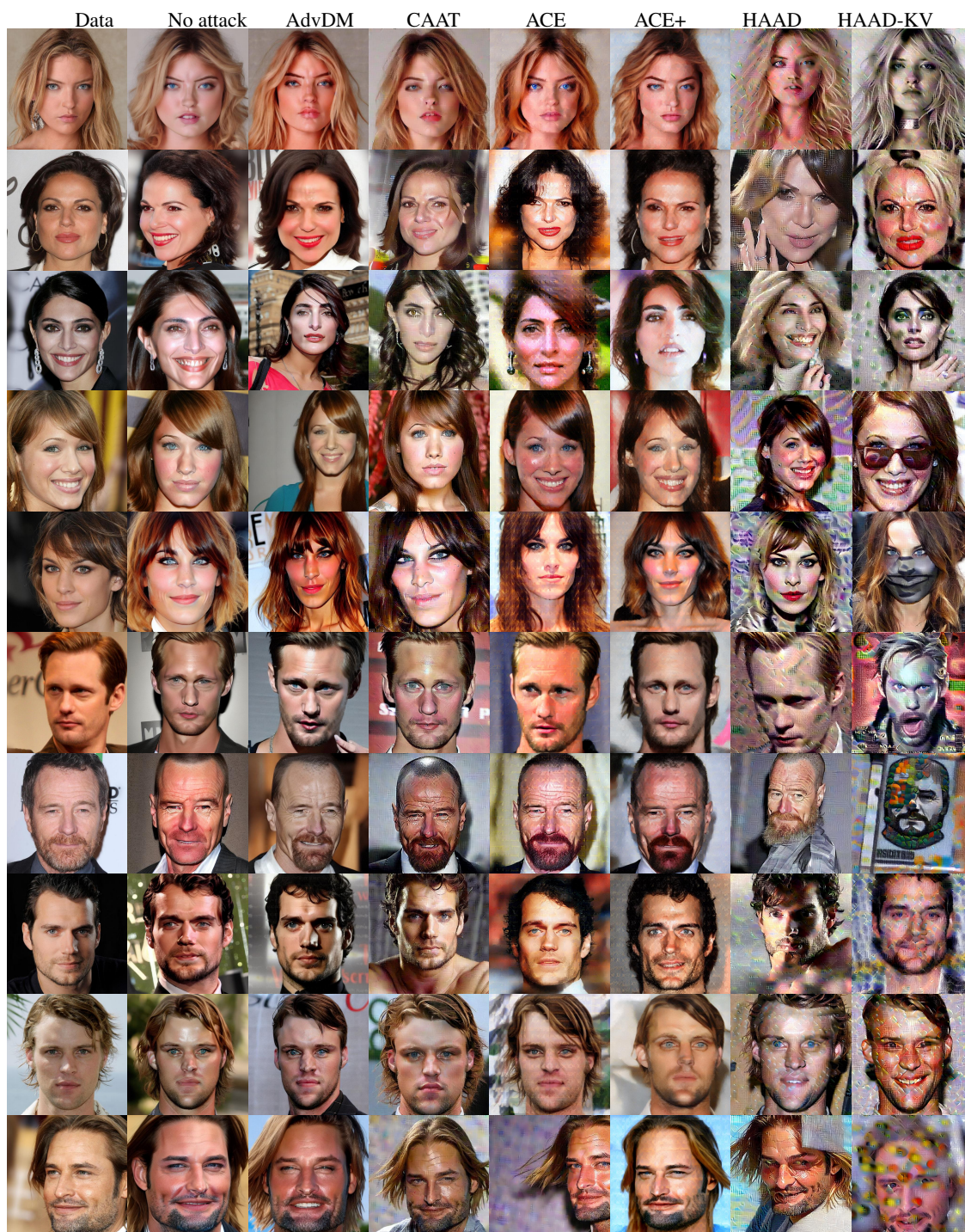




Figure 11: Comparison between different attack methods on CelebA-HQ dataset. The customization model used to generate the images is CD.



Figure 12: Comparison between different attack methods on CelebA-HQ dataset. The customization model used to generate the images is SDEdit.



Figure 13: Comparison between different attack methods on WikiArt dataset. The customization model used to generate the images is LoRA+DB.





Figure 15: Comparison between different attack methods on WikiArt dataset. The customization model used to generate the images is SDEdit.