CVPR #18418

Black Hole-Driven Identity Absorbing in Diffusion Models

Supplementary Material

In Appendix A, we delve into the Black Hole metaphor, illustrating its connection to the BIA framework and exploring identity distribution properties in high-dimensional latent space, which underpin black hole region formation. Appendix B empirically validates the boundary construction for the black hole region and provides hyper-parameter configurations for optimal efficiency. Appendix C showcases additional qualitative results, emphasizing the effectiveness of our method and some additional ablation study. Finally, Appendix D outlines the complete algorithm, offering a comprehensive overview of the BIA framework.

A. Black Hole Region Formation

This section provides a detailed explanation of the intuition behind forming the black hole region in the latent space, focusing on the efficient selection of identity-similar and dissimilar points to ensure robust boundary formation.

A.1. Black Hole Metaphor

To form a black hole, you first need stars, as black holes originate from the deaths of stars [7]. Similarly, in our problem, we draw a parallel between the latent space of diffusion models and the universe. Just as stars populate the universe, identity representations inhabit the latent space. To create a black hole in this metaphorical universe, we designate the dying identity, the one we aim to unlearn, as the black hole. This black hole absorbs nearby identity representations, gradually erasing its own identity features.

Taking inspiration from "Hubble discovers collection of black holes in globular cluster" like NGC 6397 Fig. 1, where stars are densely packed around a central black hole [3, 9], the latent space similarly exhibits clusters of identity points that are naturally close to one another. In this analogy, the surrounding identities act as neighboring stars, which are pulled toward the black hole and ultimately merged to remove the specified identity.

A.2. Identities Distribution in Latent Space

In forming the black hole region, we assume that identity-similar points are located in close proximity within the latent space. Our experiments on the CelebA-HQ dataset [2] clearly illustrate this phenomenon, as mapping 20 identities with 10 images each demonstrates that similar identities naturally cluster together within the latent space Fig. 2. This supports the notion that identity representations are inherently structured, enabling us to define a precise black hole region. By forming this region, we effectively pull in surrounding identities, facilitating the removal of the specified identity. The validity of the relationship between CelebA-HQ identities and its distribution in the latent space is also examined by GUIDE [4], which confirm the natural closeness of identities in latent space. Much like astrophysicists study the number and distribution of black holes in the universe [6], our approach models the latent space as a dynamic environment, revealing its potential for unlearning identity.

Figure 1. Globular cluster NGC 6397, captured by NASA's Hubble Space Telescope, showcasing densely packed stars surrounding a central mass of black holes. (Image credit: NASA, ESA, T. Brown, S. Casertano, and J. Anderson, STScI) [3].

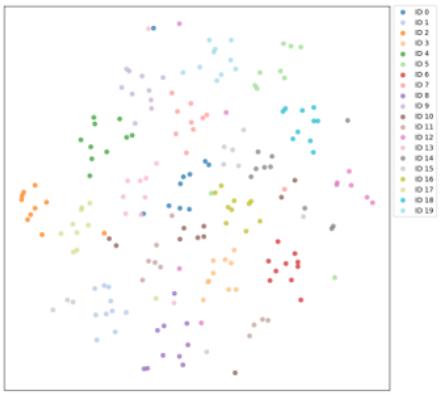


Figure 2. Latent space mapping between images and identities using t-SNE [8]. Each color represents a distinct identity with 10 images per identity from the CelebA-HQ dataset. The tight clustering of points highlights the proximity of similar identities, validating the Black Hole region's effectiveness in our BIA framework for identity unlearning.

B. Boundary Validation:

We determine the boundary in the h-space by exploring the neighborhood of the latent code h_r, representing the iden-

CVPR
#18418

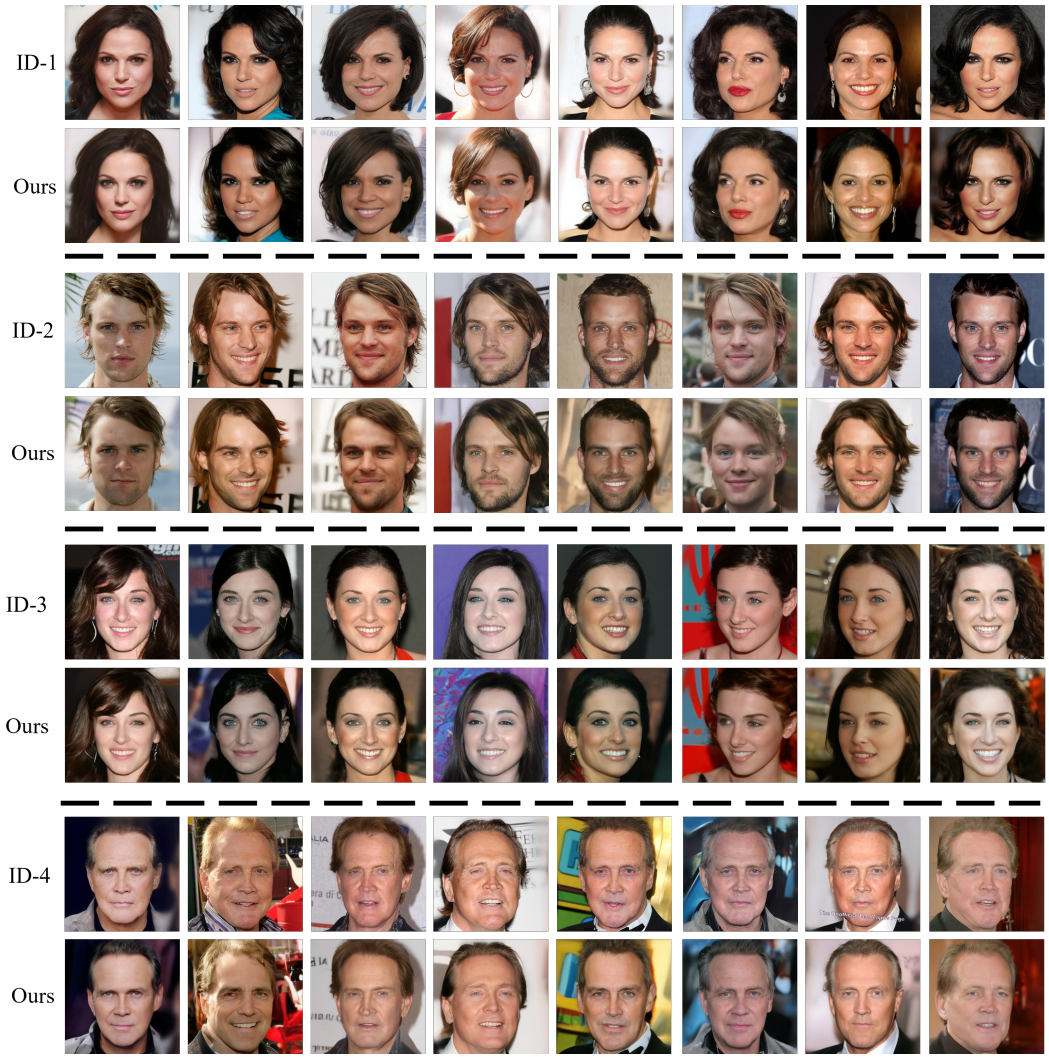


Figure 3. Additional results on the CelebA-HQ dataset showcasing different identities. The first row represents the original identity, while the subsequent row illustrates the unlearned identity achieved through our BIA framework.

tity to be unlearned, similar to [5], to ensure a balanced representation of identity-similar and dissimilar points for training a robust SVM classifier. Through experimentation with different combinations of n (number of latent points) and α (scaling factor for neighborhood exploration), we identified the optimal range of n to be between 100-150 (noted as n in the main paper, with a typo in line 262) and α between 0-20. These values effectively capture sufficient identity-similar and non-similar points in appropriate proportions for classifier training. The resulting SVM achieves over 90% test accuracy, validating the identified hyperplane d_{id} as the optimal boundary for identity disentanglement within the latent space.

C. Additional Qualitative Results:

We present supplementary qualitative results using 8 images per identity, as shown in Fig. 3. These results reaffirm that BIA effectively unlearns the target identity across both the source and other images sharing the same identity while preserving other attributes well.

In Fig. 5, we demonstrate BIA's ability to maintain the overall generative quality as outputs from the unlearned model are visually consistent with those from the original model. To ensure a fair comparison, the same seed was used to generate images before and after model unlearning. The preservation loss plays a crucial role in minimizing distribution shifts, enabling BIA to effectively unlearn the specified identity without compromising the quality of other latent codes.

CVPR
#18418

Figure 4. Unlearned images from input images with increasing value of k in Random method.

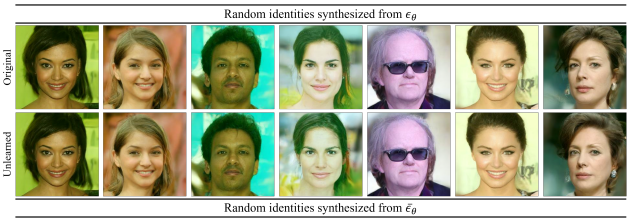


Figure 5. Qualitative comparison showcasing the consistency of BIA with the original model. The unlearned model preserves the generation quality on randomly generated images, producing nearly identical outputs to the original model.

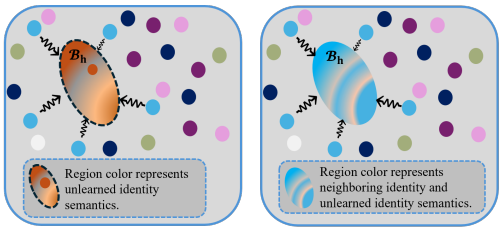
C.1. Ablation Study

In Section 4.3 of our main paper, we outlined the Random approach, which selects non-identity points to wrap the black hole region. Here, we investigate the effect of increasing the number of random points used for wrapping denoted as k_r. As k_r increases, the Random approach introduces more latent space noise, leading to a pronounced degradation in detail and structure. This progression is visualized in Fig. 4, where the first column represents the original image, and the column labeled Ours showcases the results of the BIA framework. The subsequent four columns illustrate the Random approach with increasing k_r values, demonstrating a clear decline in visual quality and detail preservation.

D. Algorithm

The latent space of diffusion models exhibits a diverse identity distribution, making it essential to define the black hole region comprehensively to encompass all latent points associated with a specified identity. To achieve this, we adopt an efficient neighborhood sampling approach inspired by GUIDE [5], but with key modifications for our framework. The overall process of our BIA framework is explained in Algorithm 1. As starting with h_r as the central reference point, we iteratively sample neighboring latent codes $\{h_j\}_{j=1}^{n}$ within its vicinity as we are only focusing on the h_r to find the neighborhood points different form the [5], which consider source and target latent both for finding neighborhood points. These neighboring points are generated by scaling the distance between h_r and randomly sampled latent codes $h$random, using a factor α drawn from a uniform distribution $\mathcal{U}(0, \alpha_{\text{max}})$, where α ranges from 0 to 20. This ensures sufficient local variation while maintaining focus on the target identity.

The sampled latent codes are categorized as identity-similar or identity-dissimilar based on their cosine similarity to x_r, evaluated using arcFace feature embeddings [1]. These labeled points are used to train an SVM, defining a hyperplane that separates identity-similar points (inside the black hole) from dissimilar points (outside). This process ensures that the black hole region comprehensively represents the target identity's latent distribution. To wrap the black hole region effectively, we incorporate $\mathcal{L}_{\text{wrapper}}$, which



a. Black hole region before unlearning b. Black hole region after unlearning

Figure 6. Illustration of Black hole region wrapping before and after unlearning.

minimizes the difference between the identity-similar and dissimilar latent points, ensuring precise identity absorption. Additionally, $\mathcal{L}_{\text{sem}}$ is used to preserve non-identity attributes during optimization, ensuring high-quality outputs. Together, these losses optimize the boundaries of the black hole region while preserving the overall semantic integrity of the latent space.

Through this process, we fine-tune the diffusion model to effectively unlearn the specified identity while maintaining the generative quality and preserving non-identity attributes as visualized in Fig. 6.

Algorithm 1: Black Hole-Driven Identity Absorption (BIA)

Input: Diffusion model ϵ_θ, unlearning identity image x_r.
Output: Updated model $\bar{\epsilon}_\theta$ with x_r id unlearned.

Initialization: Extract latent code h_r for x_r using DDIM inversion.
Step 1: Black Hole Region Formation;
for $j = 1, \ldots, n$ **do**
  Sample $h_{\text{random}} \sim \mathcal{H}$;
  Compute $\Delta_j = \alpha \frac{h_{\text{random}} - h_r}{\|h_{\text{random}} - h_r\|}$, $\alpha \sim \mathcal{U}(0, \alpha_{\max})$;
  Define $h_j = h_r + \Delta_j$;
  Compute $\text{sim}_{r,j}$ using arcFace ϕ: $\text{sim}_{r,j} = \phi(x_r) \cdot \phi(\text{Decode}(h_j))$;
  Classify h_j as identity-similar or dissimilar based on $\text{sim}_{r,j} > \text{th}_r$;
Train an SVM to compute hyperplane d_{id} separating identity-similar and dissimilar points, forming the black hole region $\mathcal{B}_h \subset \mathcal{H}$.
Step 2: Black Hole Wrapping and Identity Absorption;
Find k-nearest latent points $\{h_j\}_{j=1}^k$ (also as h_k for simplicity) around $\mathcal{B}_h$, ensuring no overlap with h_r;
Calculate wrapper loss:

$$\mathcal{L}_{\text{wrapper}} = \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{L_2}\mathcal{L}_{L_2} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}}.$$

for $h_{\mathcal{B}_h} \in \mathcal{B}_h$ **do**
  Utilize d_{id} to walk for identity-edited latent code:

$$\hat{h}_{\mathcal{B}_h} = h_{\mathcal{B}_h} + \Delta,$$

  where Δ is a step size in $[100, 120]$;
  Compute semantic loss:

$$\mathcal{L}_{\text{sem}} = \left\| h_{\mathcal{B}_h} - \hat{h}_{\mathcal{B}_h} \right\|_2.$$

Optimize:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{wrapper}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}$$

for T iterations to update ϵ_θ, yielding $\bar{\epsilon}_\theta$.
Inference Evaluate $\bar{\epsilon}_\theta$ prevents x_r generation while preserving other attributes.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3

[2] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[3] Paul Morris. Hubble discovers collection of black holes in globular cluster, 2021. 1

[4] Juwon Seo, Sung-Hoon Lee, Tae-Young Lee, Seungjun Moon, and Gyeong-Moon Park. Generative unlearning for any identity: Supplementary material. 1

[5] Juwon Seo, Sung-Hoon Lee, Tae-Young Lee, Seungjun Moon, and Gyeong-Moon Park. Generative unlearning for any identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2024. 2, 3

[6] Alex Sicilia, Andrea Lapi, Lumen Boco, Mario Spera, Ugo N Di Carlo, Michela Mapelli, Francesco Shankar, David M Alexander, Alessandro Bressan, and Luigi Danese. The black hole mass function across cosmic times. i. stellar black holes and light seed distribution. *The Astrophysical Journal*, 924(2):56, 2022. 1

[7] Paul Sutter. How many black holes are there in the universe?, 2021. 1

[8] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1

[9] Mike Wall. Scientists find clump of black holes inside the heart of globular cluster (video), 2021. 1