Latent Unlearning in Image-to-Image VQ-GAN

Gabriele Cabibbo ID: 2196717
Emanuele Gallo ID: 2197051
Giorgio Taramanni ID: 1961217

https://github.com/gheb02/Latent-Unlearning-in-Image-to-Image-VQ-GAN

Abstract

High-fidelity vector-quantized generative models such as VQ-GAN pose safety challenges due to their ability to reproduce explicit visual content. This work investigates Latent Code Replacement (LCR), a training-free unlearning strategy that operates directly on the discrete latent space by identifying and substituting codebook entries associated with unwanted content. Through a frequency-based analysis across explicit and safe datasets, we show that in high-resolution image-to-image settings explicit information is not localized in individual codes but diffused across combinations of indices. As a consequence, global code replacement strategies require large-scale substitutions and induce severe image-wide artifacts. To address this limitation, we propose Localized LCR, which constrains replacement to spatial regions detected as explicit using an external classifier. Experimental results demonstrate that localized random substitution effectively suppresses explicit features while largely preserving overall image quality, whereas semantically motivated nearest-neighbor replacements fail. Further analysis reveals that VQ-GAN codes lack independent semantic meaning and that the decoder enforces only local consistency, explaining the ineffectiveness of semantic code shifts. These findings highlight fundamental limitations of code-level unlearning in VQ-GANs and motivate future work on patch-level and co-occurrence-based latent interventions.

Warning: This research involves the analysis of explicit datasets; however, all sensitive visual content presented in this paper has been blurred or obscured to ensure the safety of the reader.

1 Introduction

High-fidelity generative models such as VQ-GAN [1] require effective safety mechanisms to prevent the generation of explicit content. Traditional unlearning approaches based on fine-tuning are computationally expensive and may degrade overall model performance. This project studies Latent Code Replacement (LCR), a training-free method that mitigates unwanted generations by identifying and substituting sensitive entries in the discrete latent space [2].
While LCR assumes that latent codes correspond to disentangled semantic concepts, we show that in high-resolution Image-to-Image (I2I) settings explicit information is diffused across combinations of codes rather than isolated entries. As a result, global replacement strategies often fail, since substituting many indices introduces image-wide artifacts, particularly affecting skin tones. To address this issue, we propose Localized LCR, which leverages NudeNet [3] to introduce spatial awareness and restrict code replacement to latent regions associated with explicit content. Our results indicate that semantic nearest-neighbor replacement is ineffective for diffused representations, whereas localized random substitution successfully suppresses explicit features while preserving overall image quality. We acknowledge that this approach does not constitute true unlearning, as it does not alter the model's internal mechanisms and can be easily worked around, similarly to applying post-hoc explicit-content masking on generated images.

2 Related Work

2.1 Vector Quantized Models

Vector Quantized (VQ) models are neural networks in which the latent space is discretized into a codebook of dimension $K \times D$, in which K is the size of the discrete latent space and D is the dimensionality of each latent embedding vector e_i. Given the output of the model's encoder, the discrete latent variable z is calculated by nearest neighbor look-up using the shared embedding space $e \in R^{K \times D}$ [4].
VQ-GAN improves upon earlier vector-quantized models by replacing pixel-wise reconstruction losses with perceptual objectives and by introducing a patch-based discriminator during training. This encourages the codebook entries to represent high-level visual structures, resulting in a compact yet expressive discrete latent space. In this way, high-resolution images are efficiently modeled as sequences of codebook indices, which can then be composed using an autoregressive transformer [1].

2.2 Latent Code Replacement

Latent Code Replacement is a training free unlearning strategy that directly operates on the codebook's discrete latent space and was first introduced in text-to-motion (T2M) models. It relies on two keys assumptions: **i.** codes

represent disentangled concepts, and **ii.** codes that represent unwanted content are identifiable [2]. If these assumptions hold, we can easily identify codes that represent unwanted contents and replace them with safer codebook entries, avoiding the generation of unwanted concepts without compromising the generation capabilities of the model.

3 Proposed Method

3.1 Explicit Codes Identification

To identify codes associated with explicit content, we utilize a frequency-based scoring method across distinct datasets. We define a Forget Set (D_f) containing more than 57000 explicit images [5] and three different Retain Sets (D_r), composed of safe pictures from ImageNet [6] and Open Images [7] (a generic subset of ImageNet, a subset of ImageNet containing images from the classes more prone of having pictures of people, and images from the "Person" class of Open Images).
Furthermore, to refine our understanding of explicit features, we employed NudeNet [3] to automatically detect and mask sensitive regions (e.g., female breasts) within the explicit dataset. This allowed us to create a specialized Masked Retain Set, where explicit body parts are occluded while preserving the surrounding safe context.
For each code k in the codebook, we calculate an activation score S_k:

$$S_k = \frac{N_k(D_f)}{N_k(D_r) + \epsilon}$$

where $N_k(D)$ denotes the frequency of index k appearing in the quantized representations of dataset D, and ϵ is a small constant to prevent division by zero.

3.2 Code Selection and Replacement Strategies

We investigate two strategies to identify and replace explicit latent codes based on activation scores S_k. For selection, we compare a threshold-based strategy ($S_k > \tau$) for variable sensitivity against a top-k strategy that enforces a fixed unlearning budget through ranking. During inference, flagged indices are replaced using either random safe replacement as a baseline or a safe nearest neighbor approach to preserve local semantic structure. As illustrated in Figure 1, global LCR primarily alters skin tones and fails to effectively suppress explicit content.

3.3 Localized LCR

As a successive refinement to the global approach, we introduce a spatial constraint to the replacement logic. By utilizing the bounding boxes provided by NudeNet, we map the pixel-level coordinates of detected explicit content to the corresponding coordinates in the latent grid.

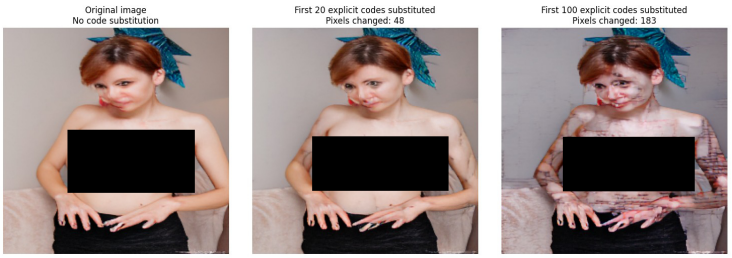


Figure 1: Global LCR primarily affects skin tone and does not remove explicit contents.

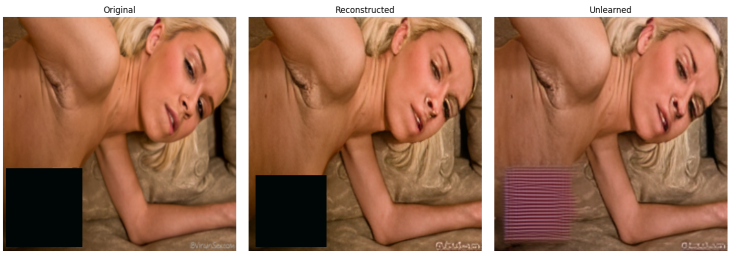
In this stage, the replacement policies are only executed if a flagged index falls within the identified explicit regions. This ensures that codes that may represent safe concepts in other parts of the image (e.g., skin tones and textures) are preserved, further minimizing the impact on the model's overall generative quality. In general, this approach also does not intervene on non-explicit images, but it has two severe criticalities: it depends on the NudeNet ability to detect explicit content and could compromise reconstruction on non-explicit images in case of False Positives.

4 Results

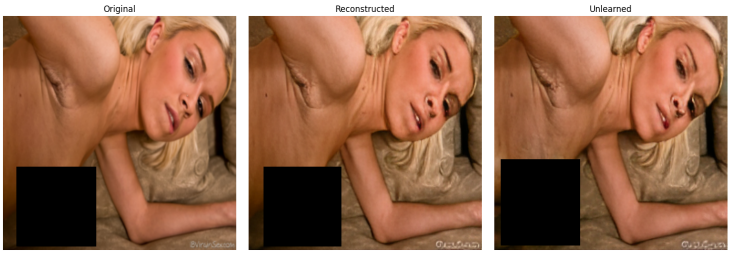
The global approaches all bring comparable results, and none of them is adequate: significant effect arise only after replacing more than ≈ 50 codes, primarily affecting skin tone but compromising the quality of the whole image (**Figure 1**). This shows that information is diffused across multiple codes, thus suggesting that individual codes may not carry semantic information.
Localized LCR shows better results: focusing on the unlearning of female breasts, we manage to remove the unwanted concept by applying a fixed, random index substitution under the detected mask (**Figure 2 (a)**), while replacing the identified explicit codes with their nearest safe semantic neighbor fails in removing the explicit concept (**Figure 2 (b)**). Moreover, we found that not all the indices located under the explicit mask are identified as explicit.
These results show that single codes are not likely to bring semantic information and, therefore, that explicit content is diffused across a combination of codebook entries.

4.1 The Shark Experiment

To further investigate the semantic properties of the codebook, we conducted an injection experiment. We extracted the latent codes corresponding to a detected explicit region from an explicit source image and injected them into the latent grid of a safe, non-explicit target image. The resulting discrete representation was then processed through the VQ-GAN decoder to evaluate the degree



(a) Reconstruction with a fixed safe index



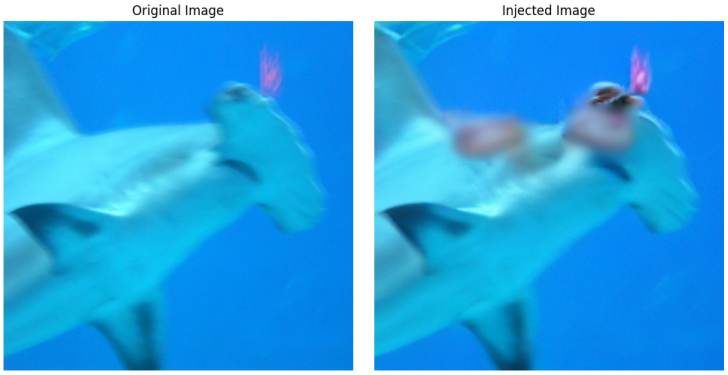
(b) Reconstruction with nearest safe semantic neighbor

Figure 2: Examples of Localized LCR

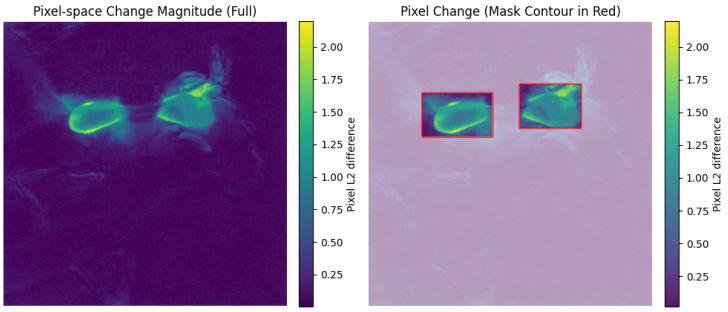
of semantic blending (**Figure 3(a)**).
If codebook entries possessed independent semantic meaning, the injected content would be expected to blend coherently with the target image's global context, adapting to variations in lighting, skin tone, or perspective. However, our observations indicate that the reconstructed features remain identical to their source appearance, lacking any contextual adaptation. Furthermore, a spatial analysis of the modified areas reveals that the changes are strictly confined to the injected coordinates and their immediate neighborhood (**Figure 3(b)**). These findings suggest that the VQ-GAN decoder operates with a localized receptive field that does not enforce global semantic consistency across the latent grid. Consequently, this lack of semantic flexibility explains why individual codes cannot be easily "shifted" toward safe neighbors and why explicit content is instead diffused across a fixed combination of entries.
These results suggest that, to perform an effective LCR-like approach in this context, we should operate on more complex structures of code entries (e.g., patches of indices or conducting an analysis on the indices co-occurrence network).

5 Conclusions and Future Work

This project evaluated the efficacy of Latent Code Replacement for unlearning explicit content in VQ-GAN models. Our results demonstrate that global code substitution is largely ineffective due to the diffused and non-semantic na-



(a) Injected image



(b) Magnitude change in the pixel space

Figure 3: Shark Experiment

ture of the discrete latent space, which leads to widespread image degradation and skin-tone artifacts. While Localized LCR provides a more targeted mitigation strategy, its success is contingent upon the accuracy of external detection frameworks like NudeNet and remains a surface-level intervention rather than a fundamental modification of the model's internal process.

The "Shark Experiment" highlights a critical limitation of the current VQ-GAN architecture: the lack of global semantic consistency and the localized receptive field of the decoder prevent codes from blending contextually. Future work should move beyond individual index substitution to explore more complex latent structures. Promising directions include the analysis of code co-occurrence networks to identify explicit "motifs" or the implementation of patch-based replacement strategies that account for the spatial dependencies of the latent grids.

Roles

- LCR with ImageNet, focused subset of ImageNet and OpenImages: Emanuele Gallo, Giorgio Taramanni
- LCR with masked retain set and Localized LCR: Gabriele Cabibbo
- Shark Experiment: Gabriele Cabibbo, Giorgio Taramanni
- Future Works (started implementing but not finished):
  - Latent Patch Replacement: Gabriele Cabibbo
  - Indices co-occurrence network and classifier counterfactual approach: Giorgio Taramanni
  - Codebook vectors cluster analysis: Emanuele Gallo

References

[1] Patrick Esser, Robin Rombach, and Björn Ommer. *Taming Transformers for High-Resolution Image Synthesis*. 2021. arXiv: 2012.09841 [cs.CV]. URL: https://arxiv.org/abs/2012.09841.

[2] Edoardo De Matteis et al. *Human Motion Unlearning*. 2025. arXiv: 2503.18674 [cs.CV]. URL: https://arxiv.org/abs/2503.18674.

[3] notAI-tech. *NudeNet: lightweight Nudity detection*. https://github.com/notAI-tech/NudeNet. 2024.

[4] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. *Neural Discrete Representation Learning*. 2018. arXiv: 1711.00937 [cs.LG]. URL: https://arxiv.org/abs/1711.00937.

[5] alex000kim. *nsfw_data_scraper*. https://github.com/alex000kim/nsfw_data_scraper. 2019.

[6] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[7] Alina Kuznetsova et al. "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale". In: *International Journal of Computer Vision* 128.7 (Mar. 2020), pp. 1956–1981. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01316-z. URL: http://dx.doi.org/10.1007/s11263-020-01316-z.