Memories of Forgotten Concepts
(CVPR 2025)

Previous paper

'**Concept Ablation** in Diffusion Models'

[1] Erasing Concepts from Diffusion Models (CVPR 2023)

*"We propose a fine-tuning method that can **erase a visual concept** from a pre-trained diffusion model, given only the name of the style and using negative guidance as a teacher."*

[2] Erasing Undesirable Concepts in Diffusion Models with Adversarial Preservation (NeurIPS 2024)

*"we propose to identify and preserving concepts most affected by parameter changes, termed as adversarial concepts. This approach ensures stable **erasure with minimal impact on the other concepts**."*

Memories of Forgotten Concepts (CVPR 2025)

*"we reveal that the **erased concept information persists** in the model and that erased concept images can be generated using the right latent."*

[Hypothesis]

"Hypothesis: An ablated model **should not have a high likelihood seed vector** that can be used to generate a high-quality ablated image."

[Method]

"**Utilizing inversion methods**, we show that there exist latent seeds capable of generating high quality images of erased concepts."

"we … **analyze the likelihood of the corresponding seed** … as well as the **quality of the generated image**."

[Results]

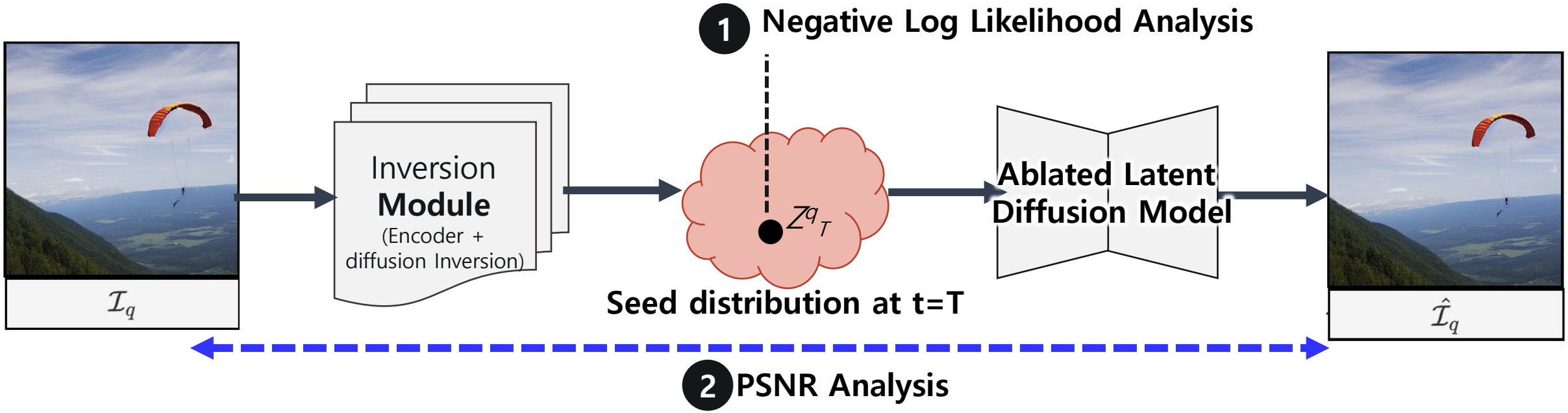
"at the dataset level, **there exists at least one latent per image that can reconstruct the image** with high quality (PSNR ≥ 25 dB) from a reasonable likelihood."

[Conclusion]

"… our analysis shows that **the opposite holds true**."

Experimental Setup

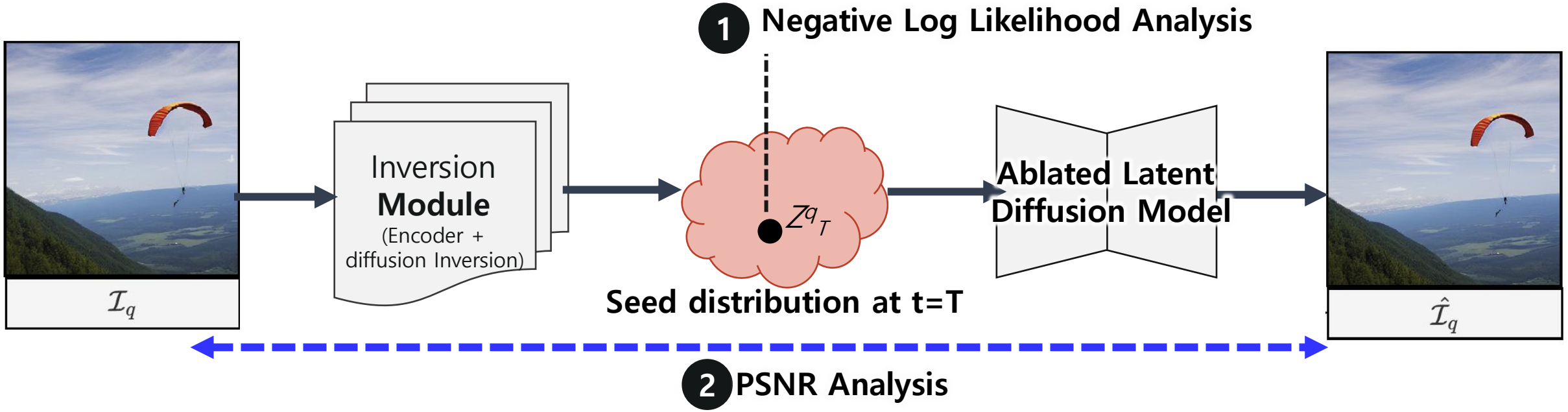
I_q : ablated query image
z_T : latent seed vector (t=T).
$z_T{}^q$: retrieve a latent seed vector
$\hat{I}_q$: reconstructed image from $z_T{}^q$



(Simplified) Figure 4. Memory of an ablated image:
Given an ablated query image I_q, our goal is to find a likely latent z_T that can accurately reconstruct the image when processed through an ablated diffusion model.
We start by encoding I_q into a latent z_0 with the encoder, then apply diffusion inversion to obtain a seed latent vector z_T . This seed is fed into the LDM to generate the image $\hat{I}_q$. Finally, we evaluate the likelihood of z_T and the quality of the reconstructed image $\hat{I}_q$ compared to I_q.

Experimental Setup

I_q : ablated query image
z_T : latent seed vector (t=T).
$z_T{}^q$: retrieve a latent seed vector
$\hat{I}_q$: reconstructed image from $z_T{}^q$



1 Negative Log Likelihood Analysis

Instance level

$$\mathrm{NLL}(Z) = \frac{k}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{k}(Z_i - \mu)^2. \quad (9)$$

Dataset level

$$d_{\mathcal{N}}(E, R) := \frac{\mathrm{EMD}\left(\mathrm{NLL}_{\to z_T}(E), \mathrm{NLL}(\mathcal{N})\right)}{\mathrm{EMD}\left(\mathrm{NLL}_{\to z_T}(R), \mathrm{NLL}(\mathcal{N})\right)}. \quad (5)$$

E (erased set) : pairs with images containing concept c.
R (reference set) : pairs without images containing concept c.
EMD :earth mover's distance

2 PSNR Analysis

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

MAX_I : maximum possible pixel value of the image (wikipedia)

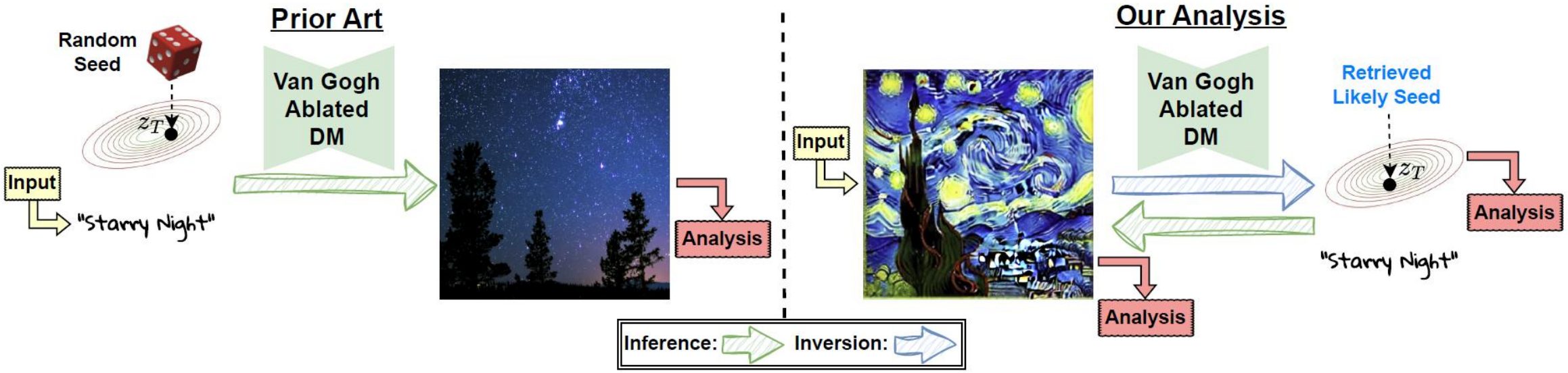


Figure 1. Evaluation of concept erasure models: Prior Art vs. Our Analysis.
Prior art analyzes the image generated by an ablated model using the text (or textual embeddings) and a random seed. Instead, we assume that both text and ablated image are given and analyze the likelihood of the corresponding seed, in the latent space of the model, as well as the quality of the generated image. We find that ablated models contain seeds with high likelihood that can be used to generate high quality ablated images

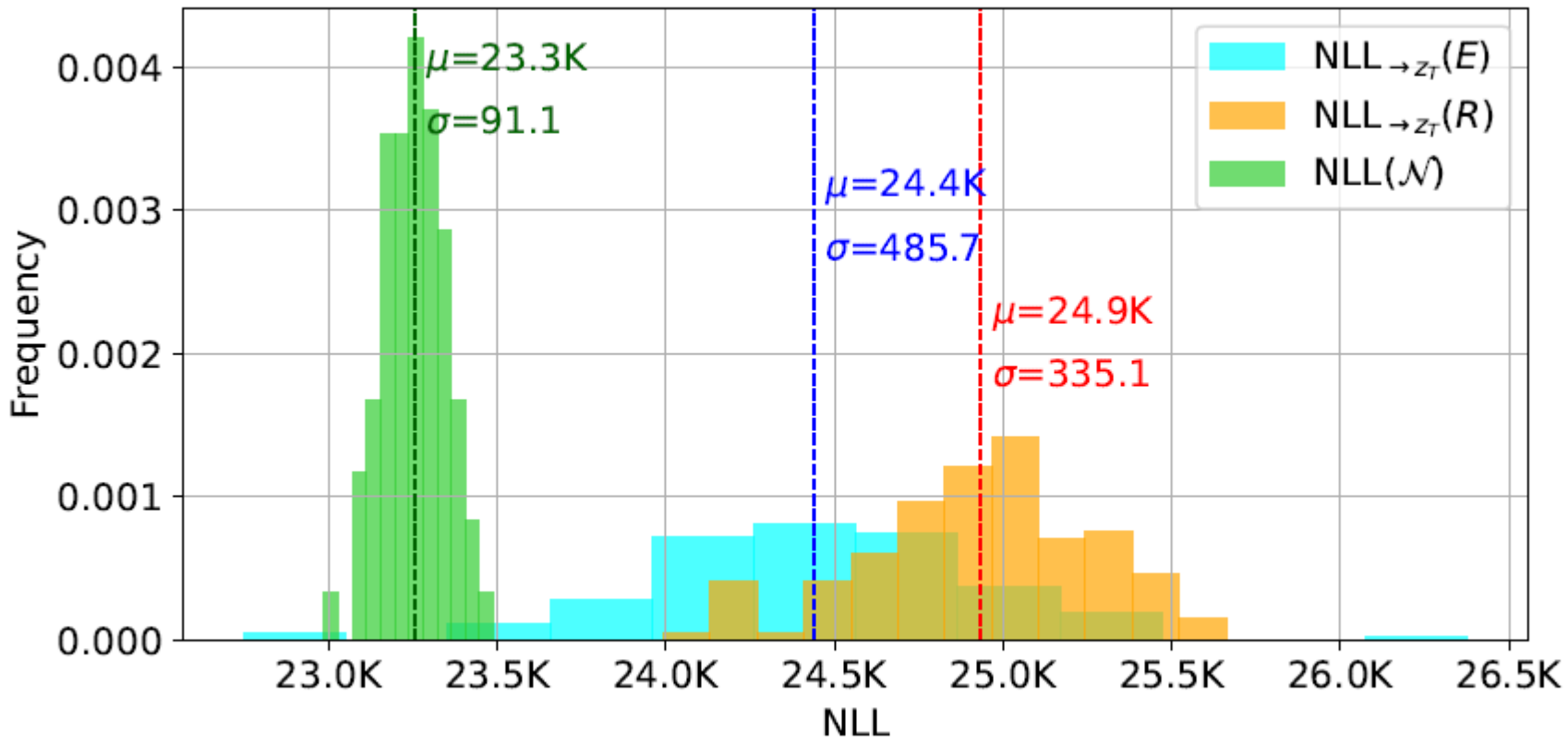


Figure 2. NLL histogram: For a model that erased the concept Nudity (EraseDiff [43]), the likelihood distribution fits different Gaussians (NLL→zT(E), NLL→zT(R)), that are different from the sampling distribution of the LDM which is standard normal distribution (NLL(N)).

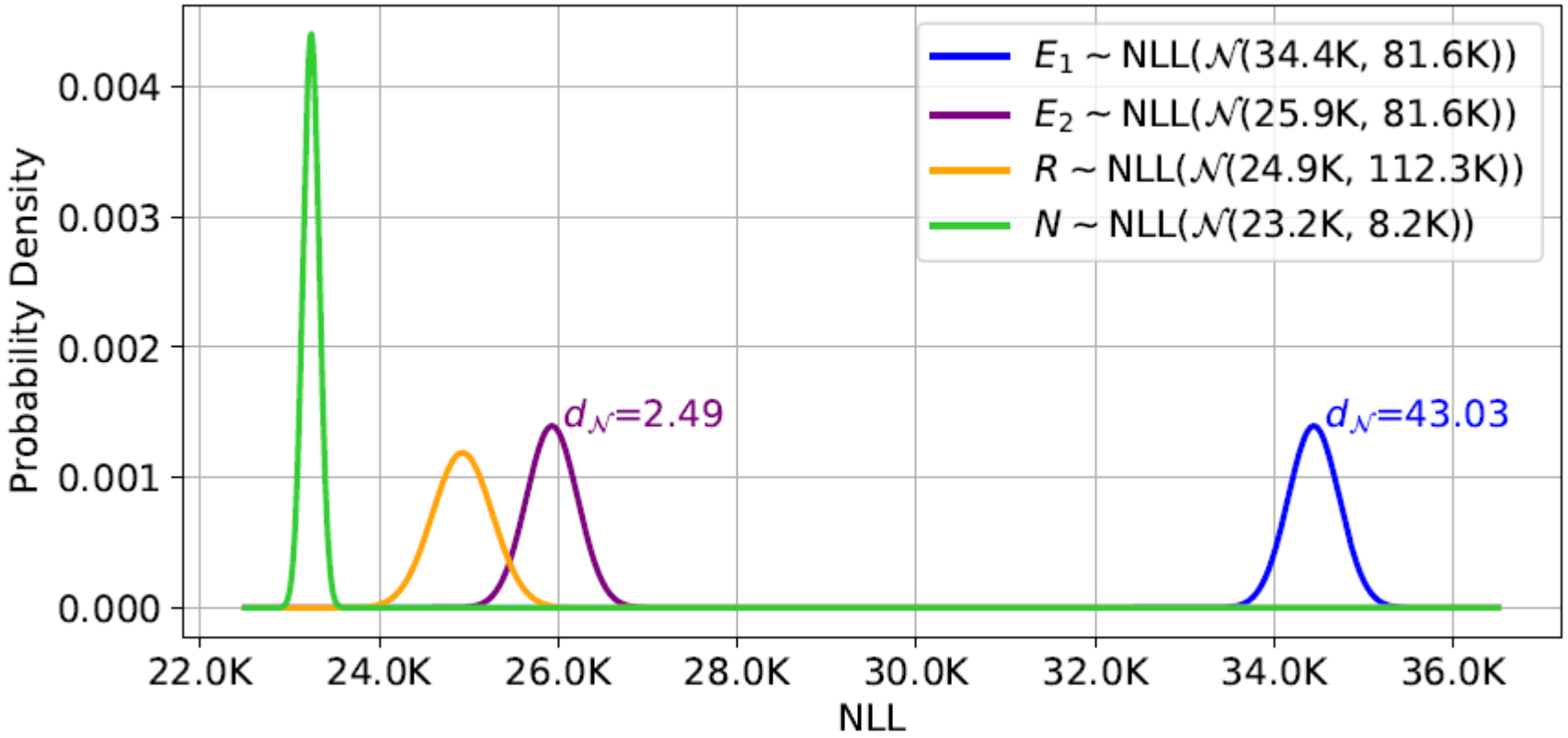


Figure 3. Visualizing our distance measure: Our relative distance measure is the ratio of EMD(E,N)) to EMD(R,N)), where E is the erased set, R is the reference set, N is the normal distribution, and EMD is Earth Movers Distance. As can be seen, the erased model E1 is much farther than E2, suggesting that the model that forgot E1 did a much better job.

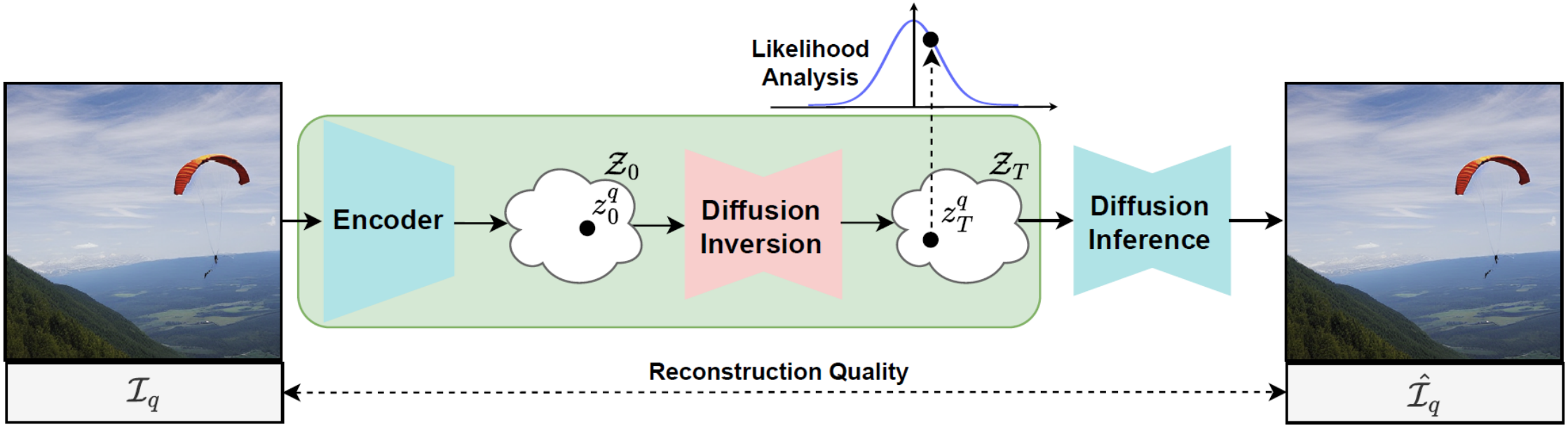
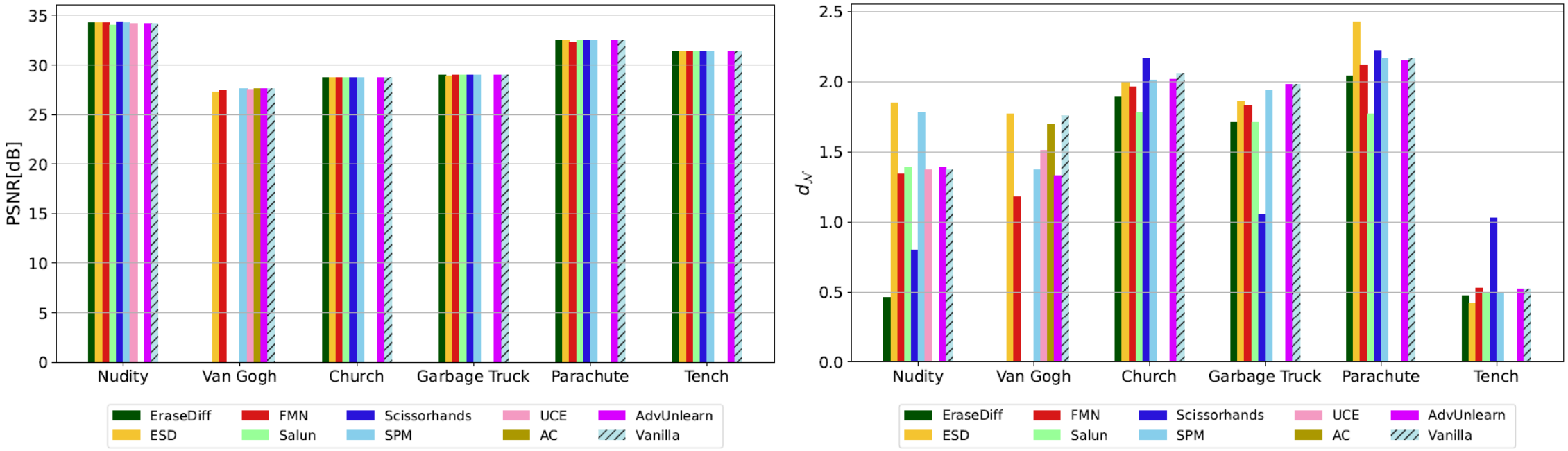


Figure 4. Memory of an ablated image:
Given an ablated query image I_q, our goal is to find a likely latent z_T that can accurately reconstruct the image when processed through an ablated diffusion model.
We start by encoding I_q into a latent z_0 with the encoder, then apply diffusion inversion to obtain a seed latent vector z_T . This seed is fed into the LDM to generate the image $\hat{I}_q$. Finally, we evaluate the likelihood of z_T and the quality of the reconstructed image $\hat{I}_q$ compared to I_q.



(a) PSNR [dB]

(b) $d_{\mathcal{N}}(E, R)$

Figure 5. A concept erased model remembers: We report the mean reconstruction PSNR (a) and our proposed relative distance (b)
for six concept datasets {Nudity, Van Gogh, Church, Garbage Truck, Parachute, Tench} across nine different concept ablation methods
{EraseDiff [43], ESD [7], FMN [45], Salun [4], Scissorhands [42], SPM [22], UCE [8], AC [18], AdvUnlearn [47]}, along with one
"Vanilla" SD 1.4 [30] model. These results validate that, at the dataset level, there exists at least one latent per image that can reconstruct
the image with high quality (PSNR ≥ 25 dB) from a reasonable likelihood using the concept erased model.

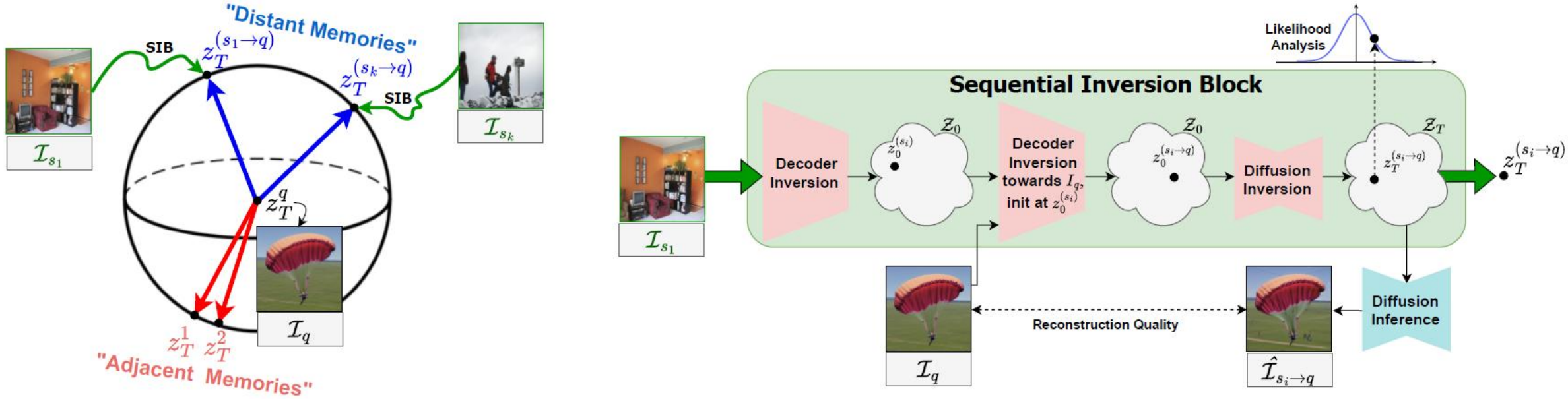


Figure 6. The many memories of an ablated image: (Left) The latent seed zq
T of the query image Iq can be obtained from various support
images: Is1 , ..., Isk . For support image Isi , we apply the sequential inversion block shown on the right to map it to the seed z(si→q)
T . We
show in Fig. 5 that seeds {z(si→q)
T }ki
=1 are likely enough and can be used to generate the query image Iq. (Right) Recovering the seed of a query image Iq when starting with support image Isi .

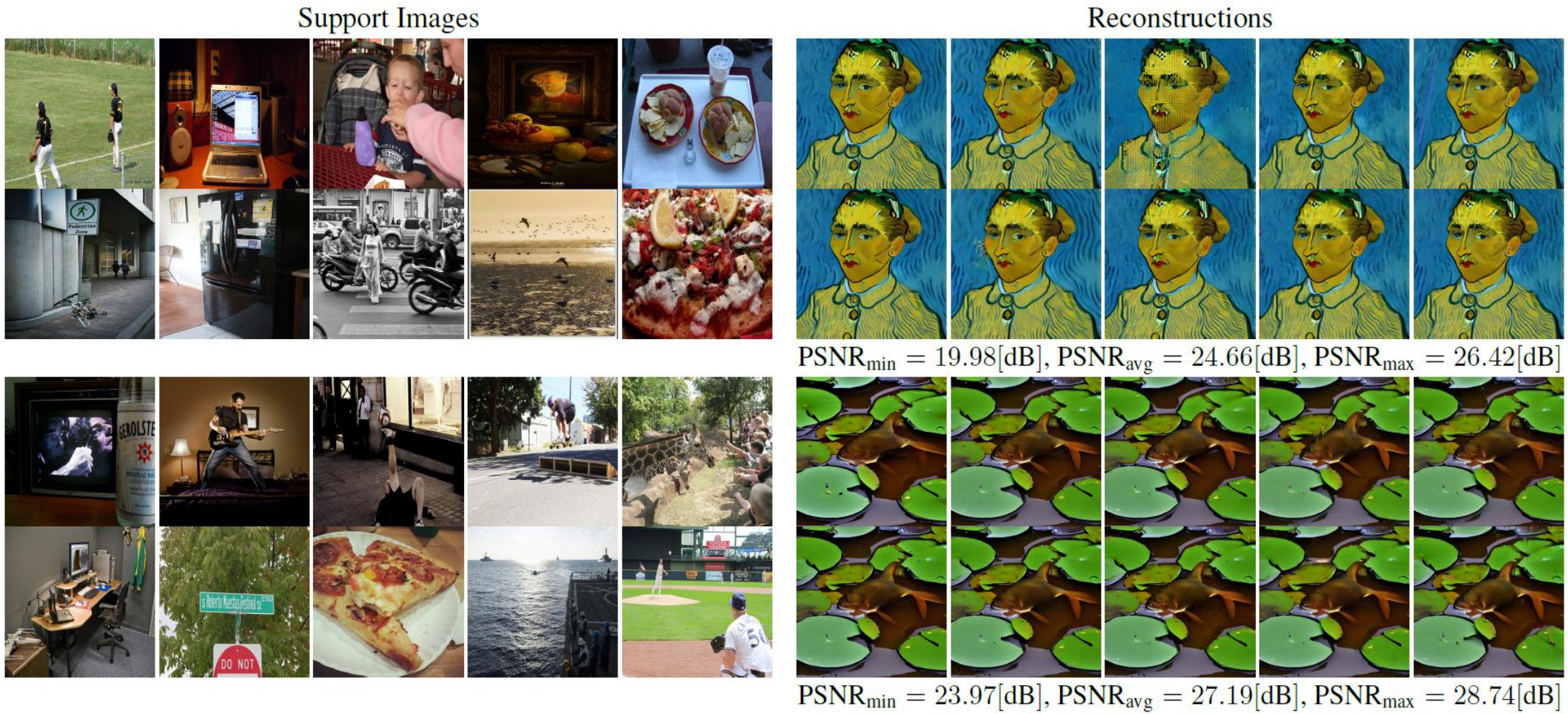
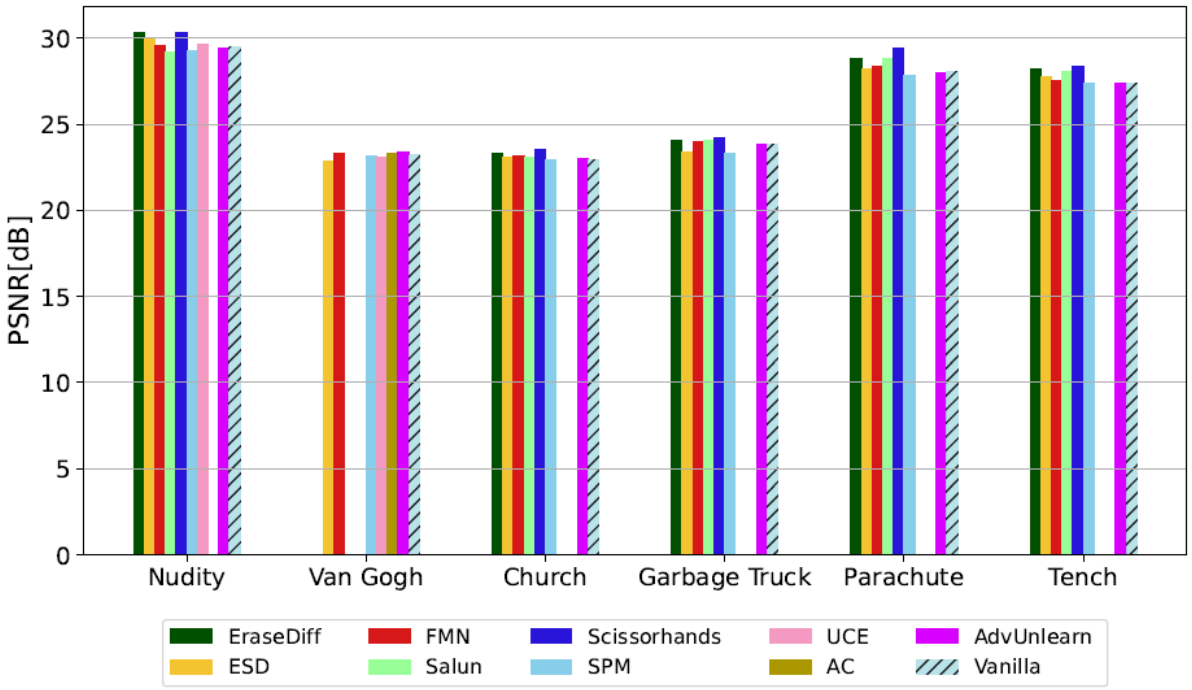
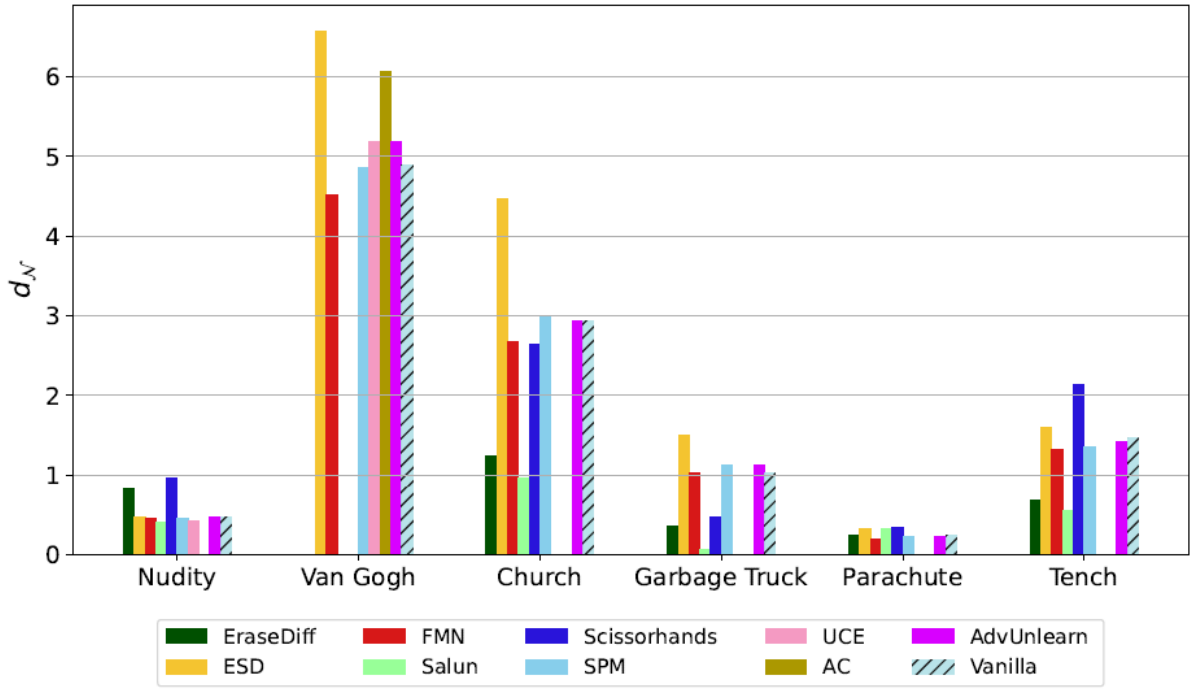


Figure 7. Reconstruction from different seeds: Reconstructions from multiple zT seeds were generated via the Sequential Inversion Block (see Sec. 3.3) using 10 different support images (left) for two target concepts: Van Gogh and Tench. While preserving each concept's core appearance, reconstructions vary slightly in background texture, blurriness, and minor elements (e.g., shirt buttons, fins). The average cosine distances between seeds are 0.58 for Van Gogh and 0.69 for Tench. These images were generated from an ESD [7] model that ablated the Van Gogh (Top) and Tench (Bottom) concepts.



(a) PSNR [dB]



(b) $d_{\mathcal{N}}(E, R)$

Figure 8. Distant latents reconstruct erased images: We report the mean reconstruction PSNR (a) and our proposed relative distance (b), for different models and concepts (see Fig. 5 for more details), obtained using our sequential inversion block process resulting in different distant latents for each image.

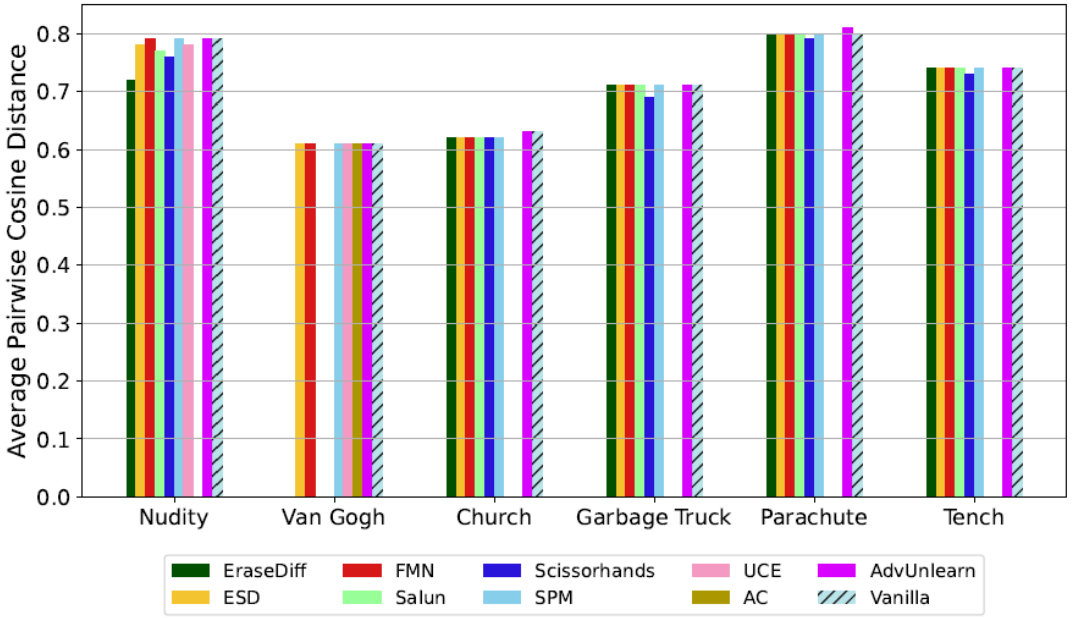


Figure 9. Average Pairwise Cosine Distance: For each model that ablated each concept, and for each target image Iq, we average
the pairwise cosine distance (1−cosine similarity) between all the produced z(si→q)
T seed latents. Then, we average the results
over all target images per each model and concept.

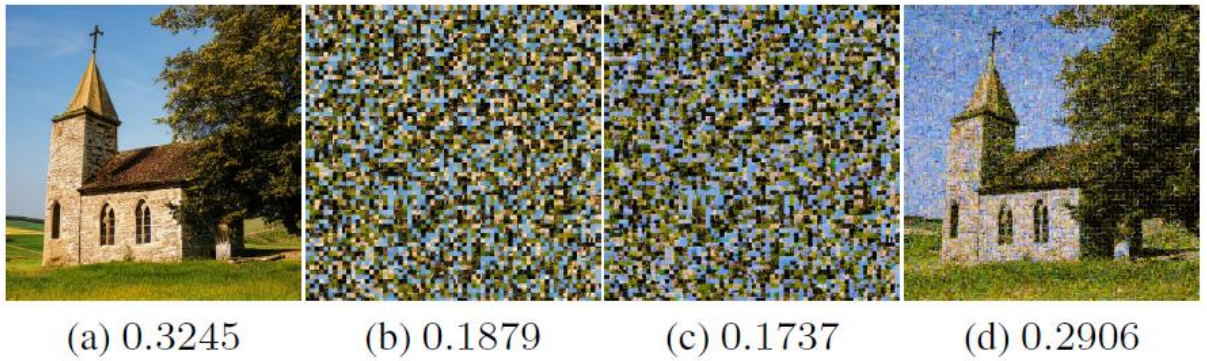


Figure 10. Ablated models generalize to shuffled images: For a diffusion model that ablates the concept Church, we take a church image in (a), split it to patches of shape 8 × 8 and shuffle them to obtain image (b). Then, we invert the image in (b) and regenerate it to obtain the reconstructed image in (c). Finally, we revert the shuffle of patches to obtain the image at (d). Below each image we report the CLIP score of that image w.r.t. the text "church".

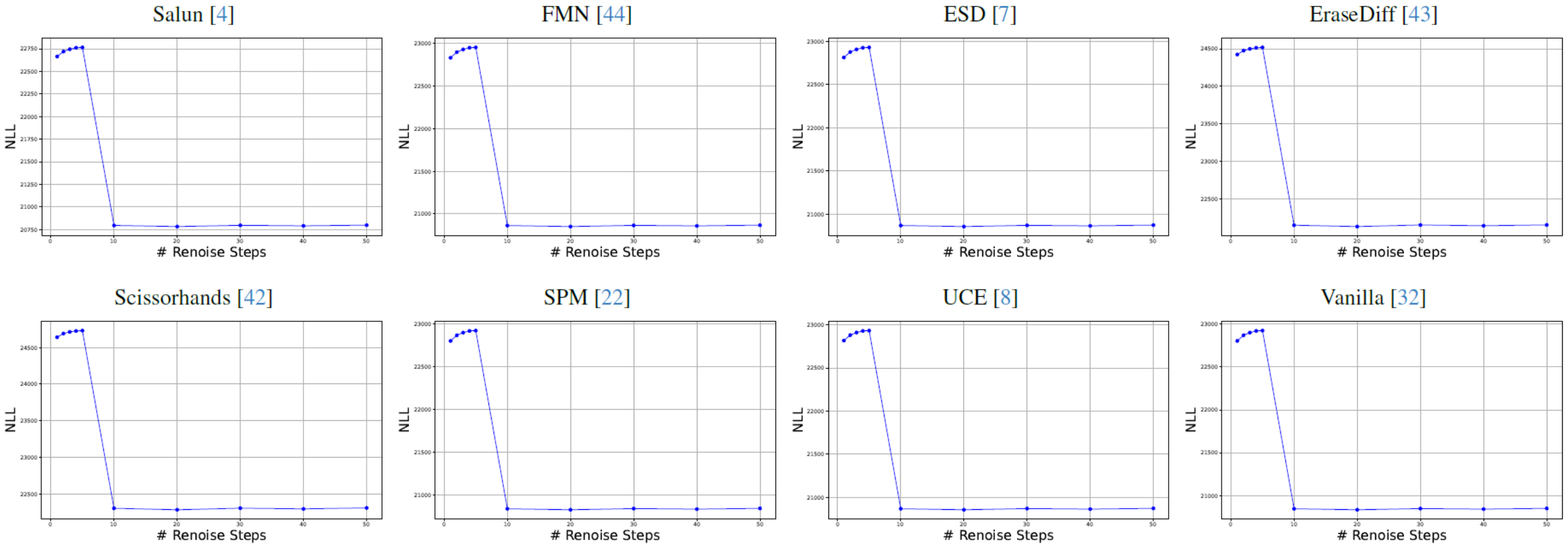


Figure 16. Choosing the right renoising parameter. Using Renoise [9], we see that after a certain amount of iterations, the NLL drops dramatically, making it harder to perform a likelihood analysis.