Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models

Gong Zhang[1*] , Kai Wang[1*] , Xingqian Xu[1,3], Zhangyang Wang[2,3] , Humphrey Shi[1,3]
[1] SHI Labs @ Georgia Tech & UIUC, [2] UT Austin, [3] Picsart AI Research (PAIR)

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition **(CVPR) Workshops, 2024**
Citations: 189 (as of April 02, 2025)

Presenters

Subhajit Maity
Nakul Rathore
Srinivas Venkatanarayanan
Jaiminkumar Ashokbhai Bhoi

Advisor

Dr. Mubarak Shah

Disclaimer: This presentation contains redacted NSFW images.

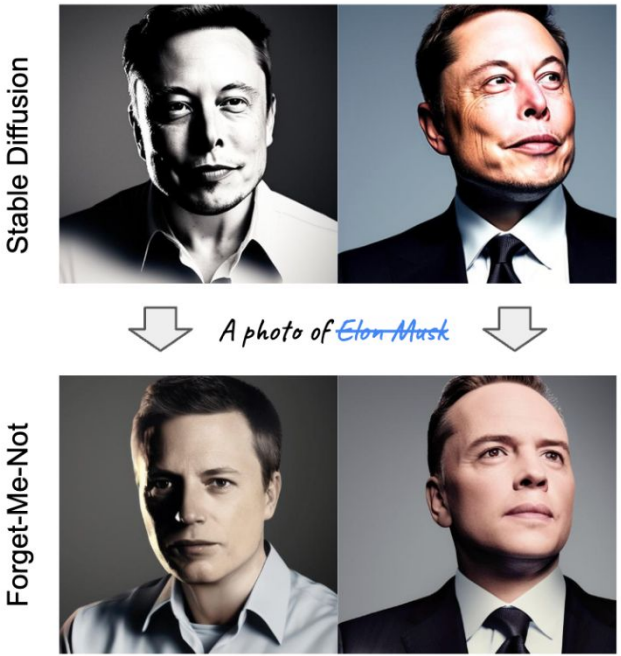
Problem - Forgetting a Concept

Stable Diffusion

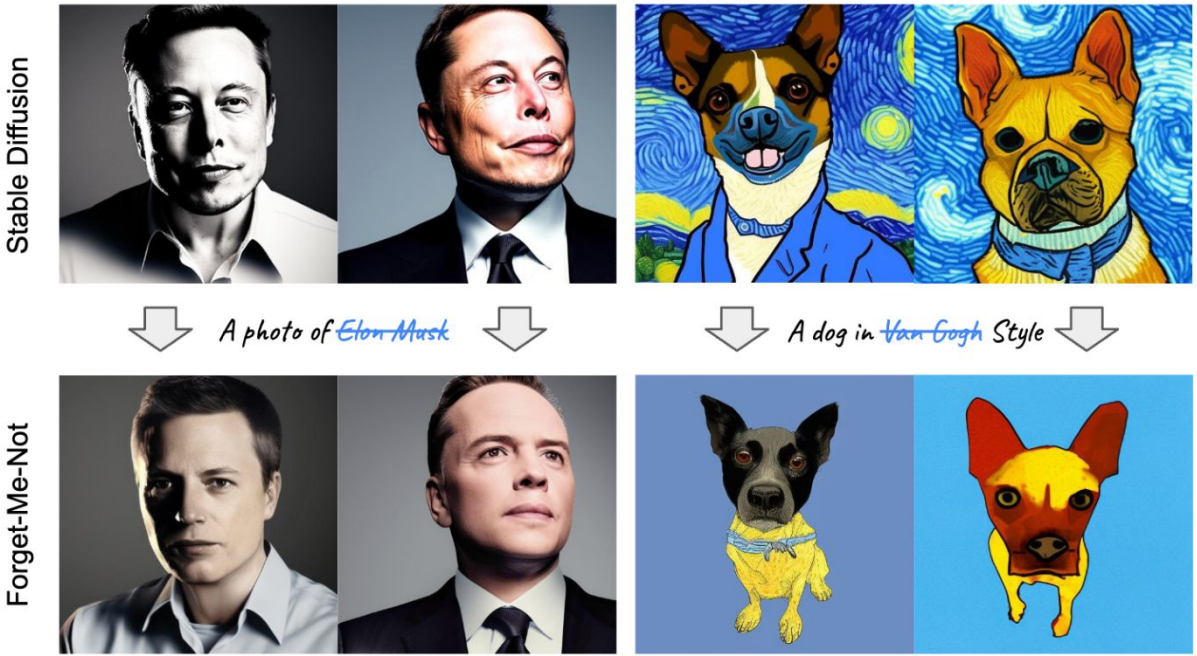


A photo of Elon Musk

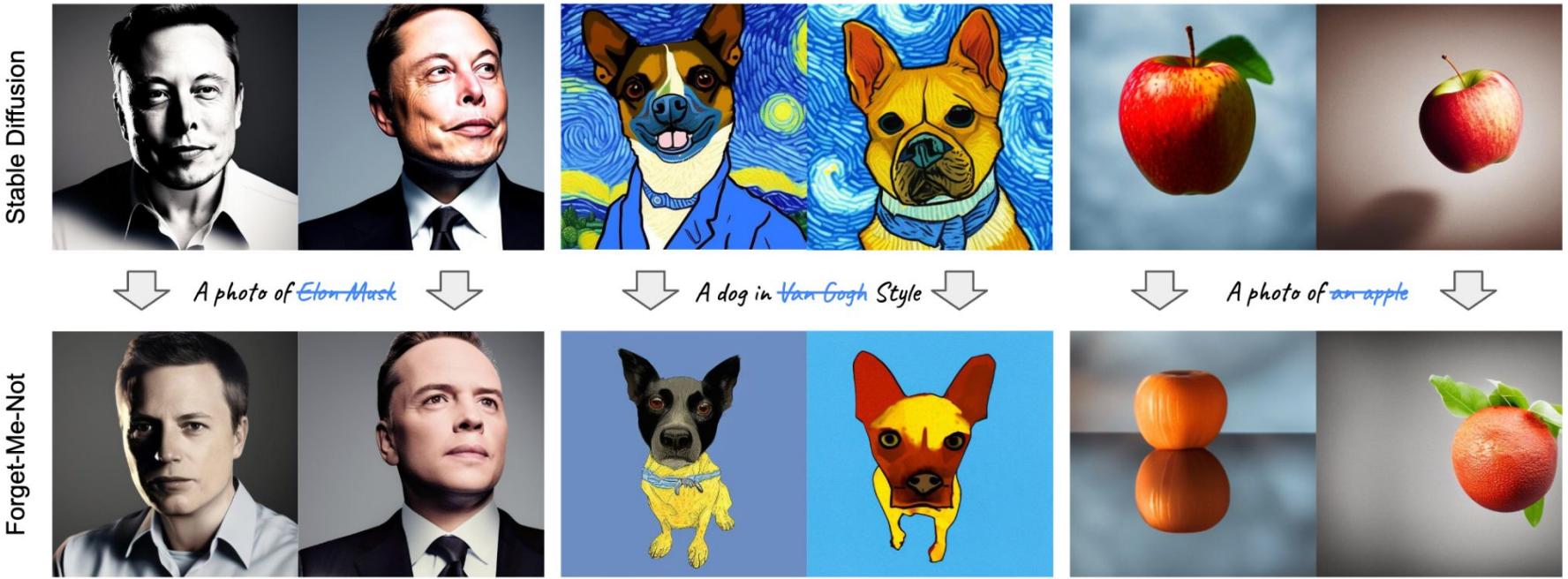
Problem - Forgetting a Concept



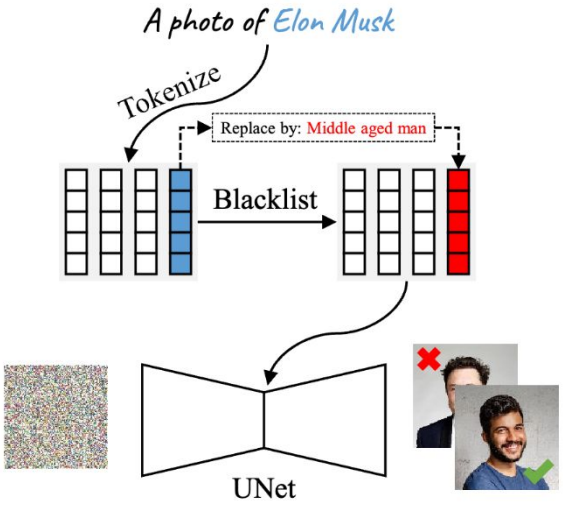
Problem - Forgetting a Concept



Problem - Forgetting a Concept

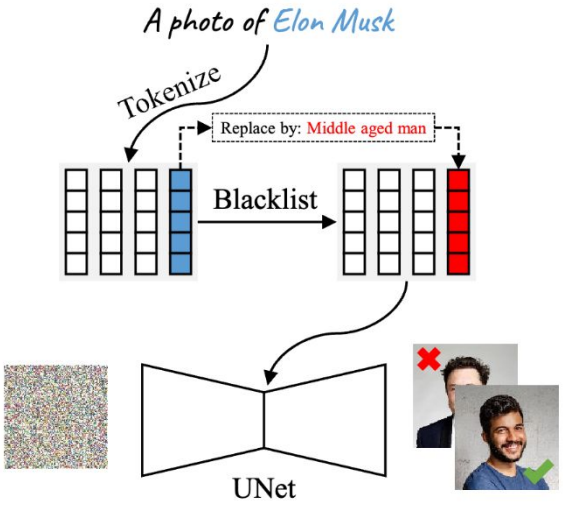


Forgetting Methods



(a) Token Blacklist

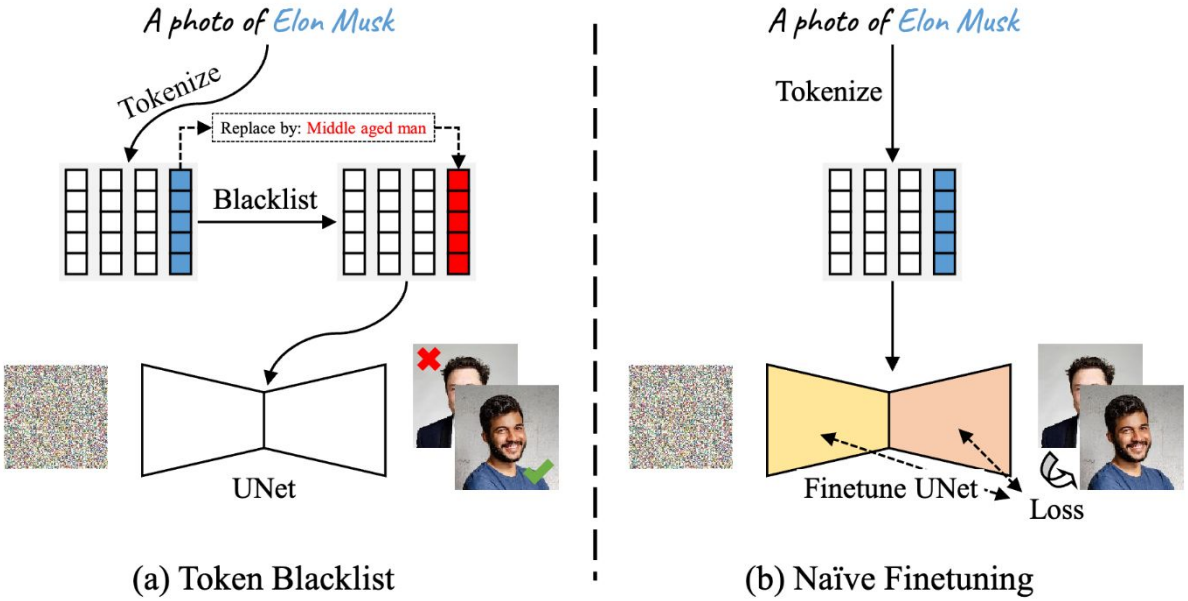
Forgetting Methods



(a) Token Blacklist

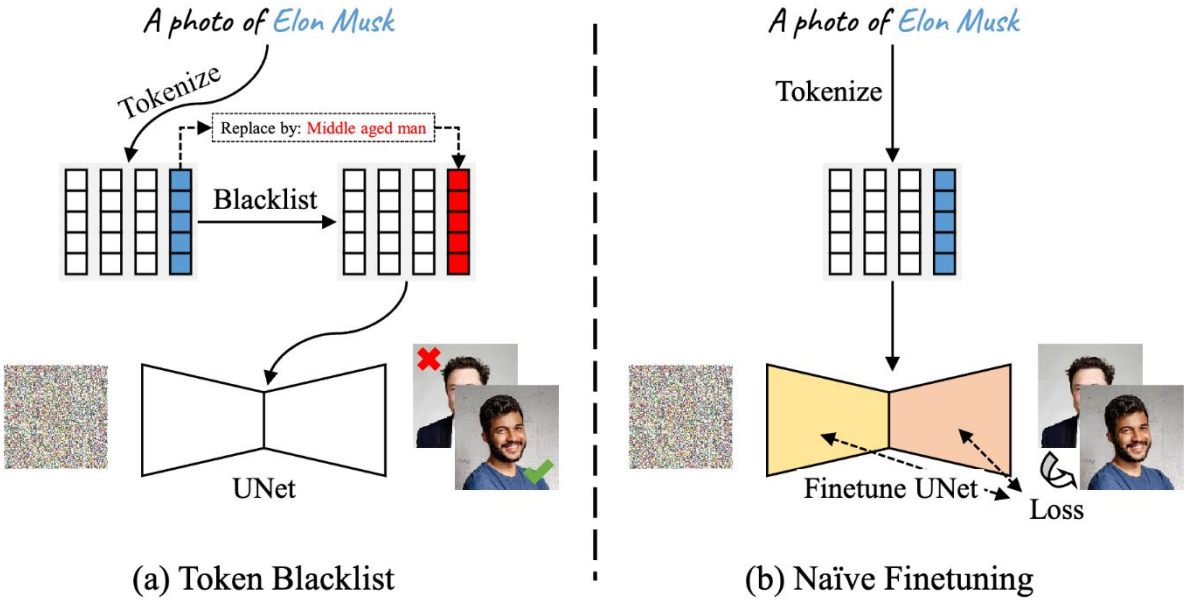
No forgetting

Forgetting Methods



No forgetting

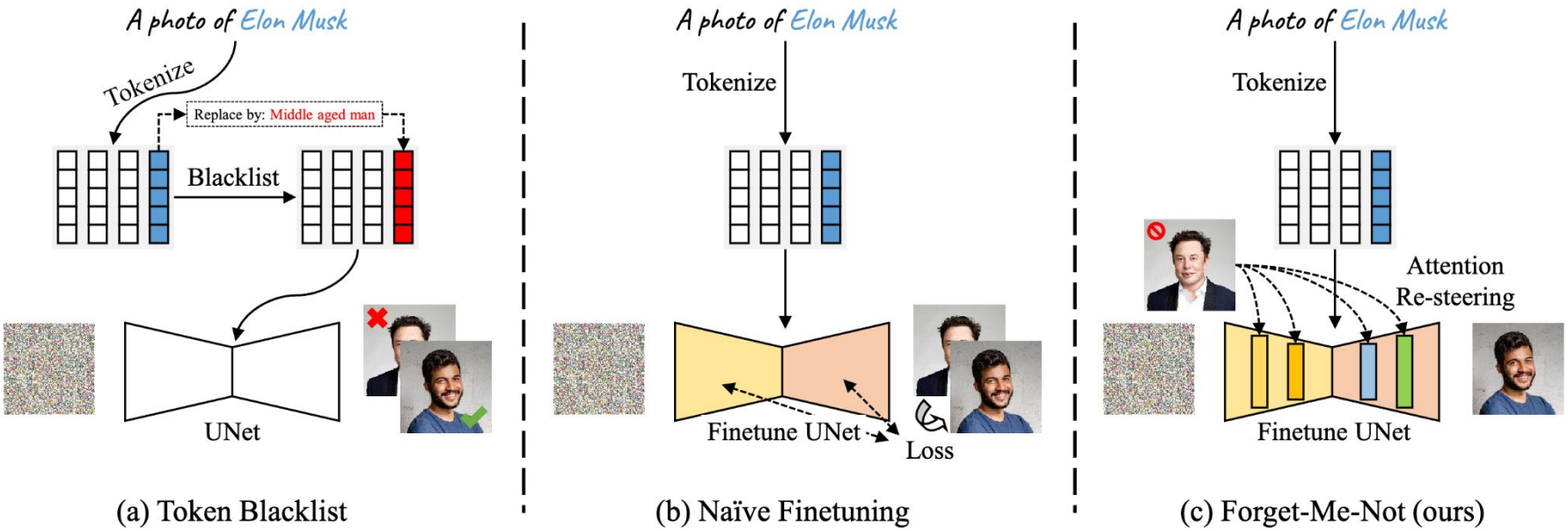
Forgetting Methods



No forgetting

Impacts other concepts

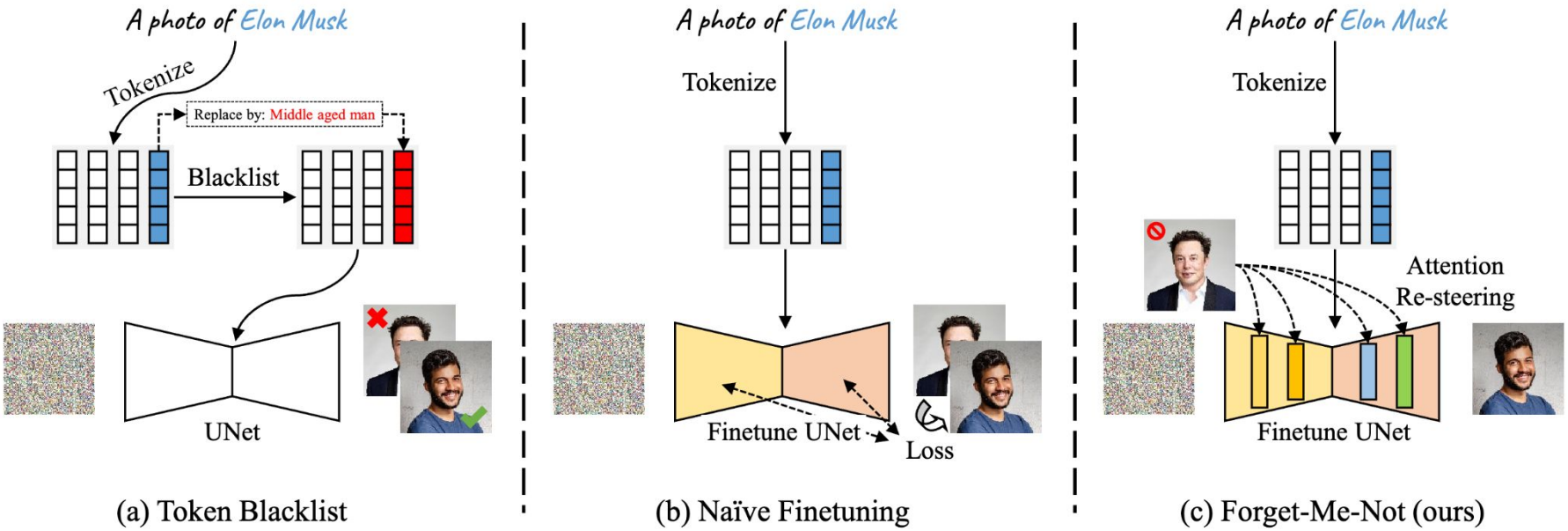
Forgetting Methods



No forgetting

Impacts other concepts

Forgetting Methods



(a) Token Blacklist

(b) Naïve Finetuning

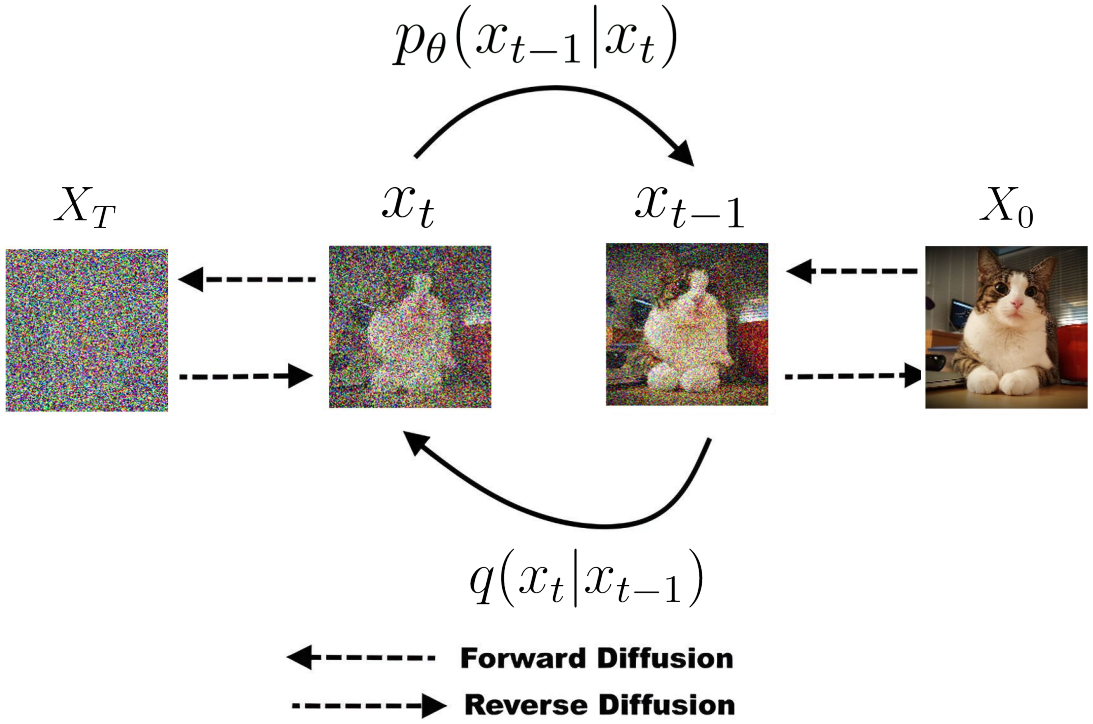
(c) Forget-Me-Not (ours)

No forgetting

Impacts other concepts

Cross-Attention based technique

Diffusion Model



Diffusion Model Equations

Forward process

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$$

Diffusion Model Equations

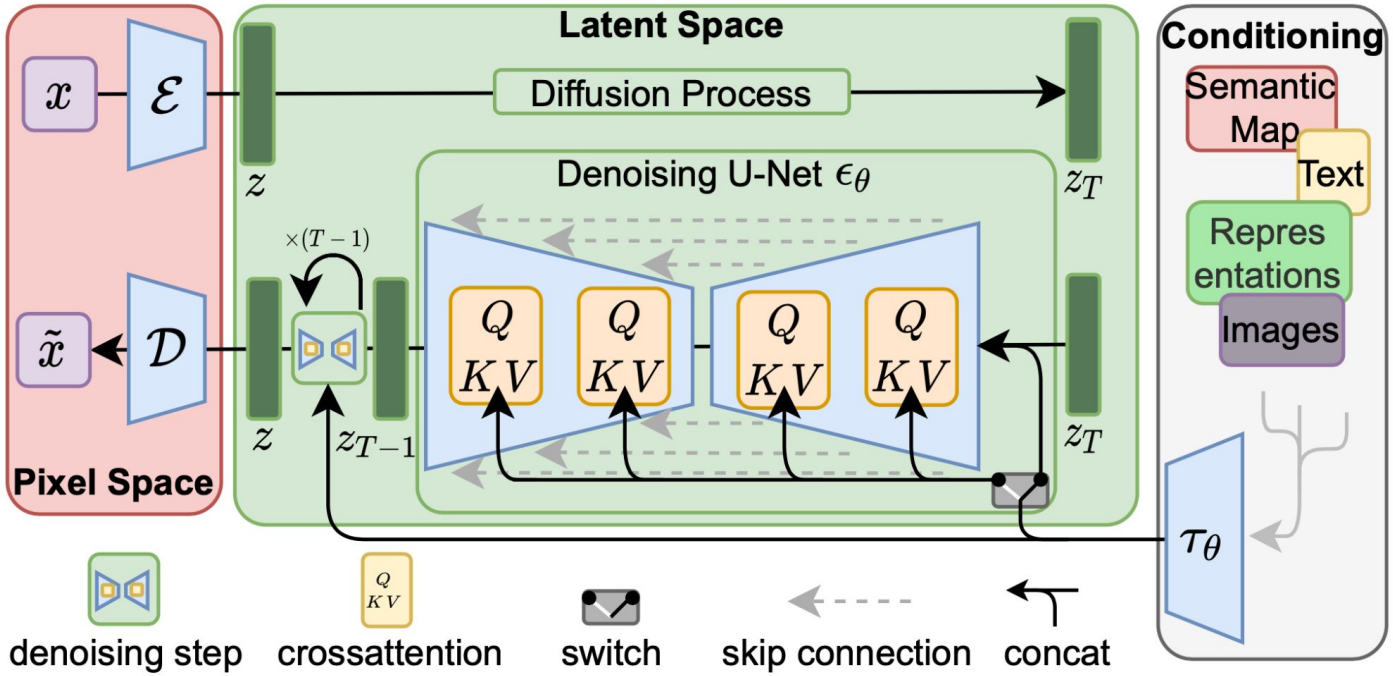
Forward process

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$$

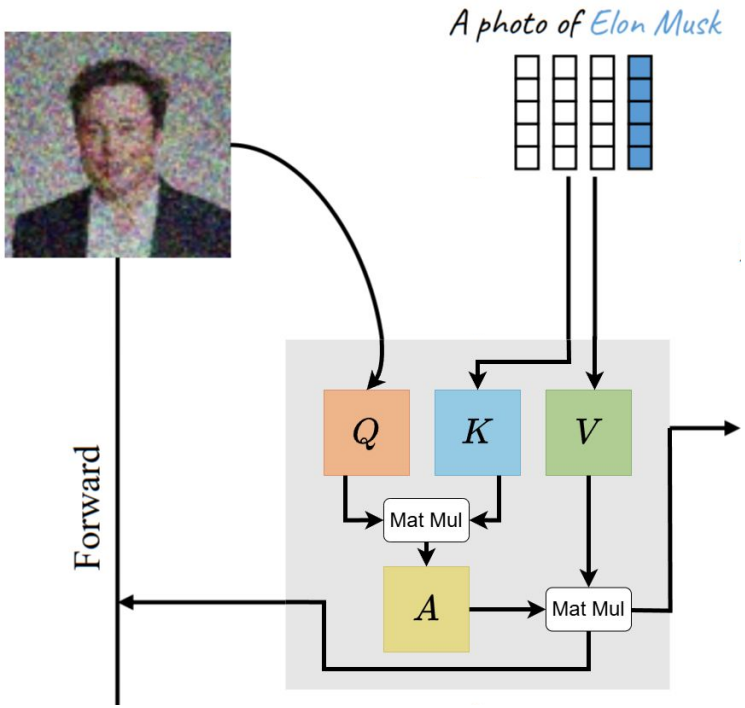
Reverse process

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

Diffusion Model Architecture

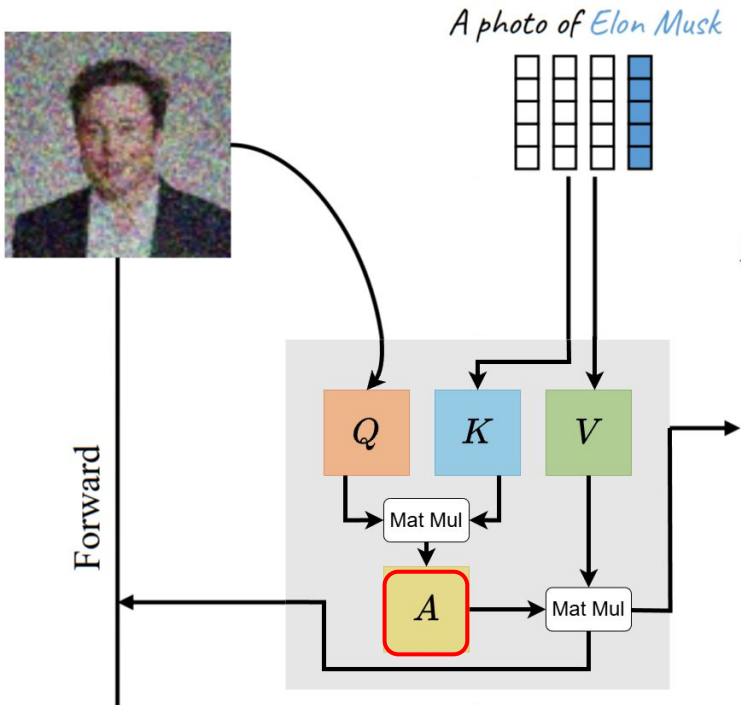


Cross-Attention



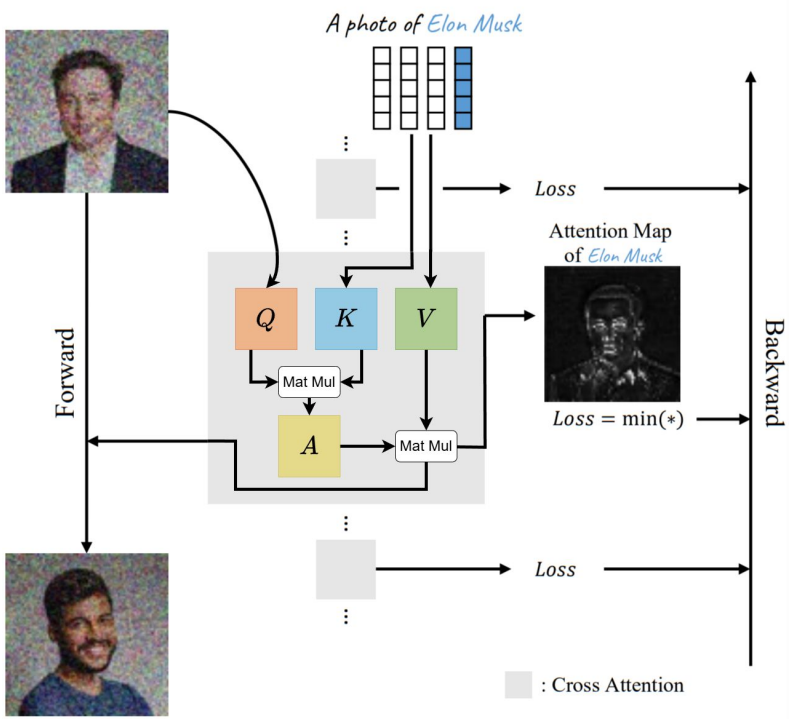
$$h = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

Cross-Attention



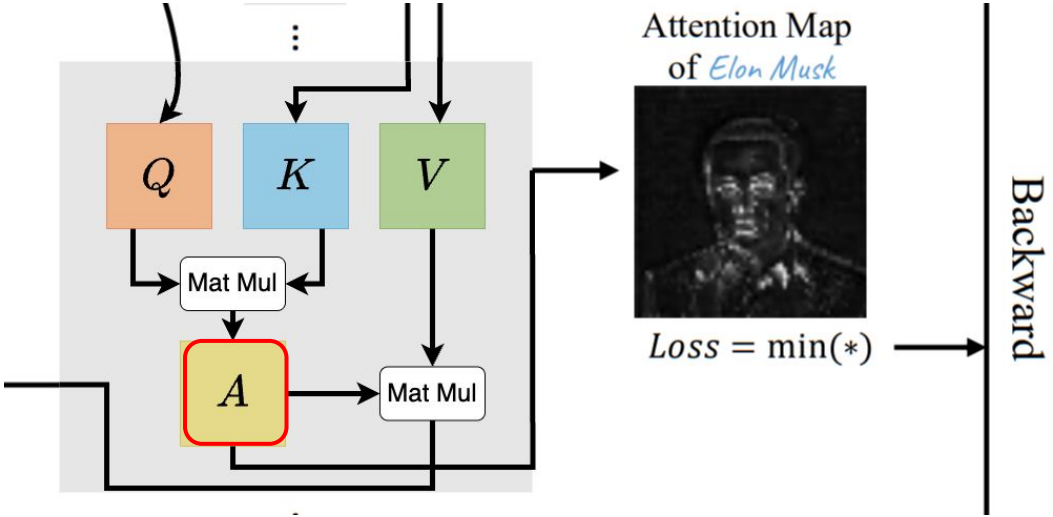
$$h = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

Cross-Attention



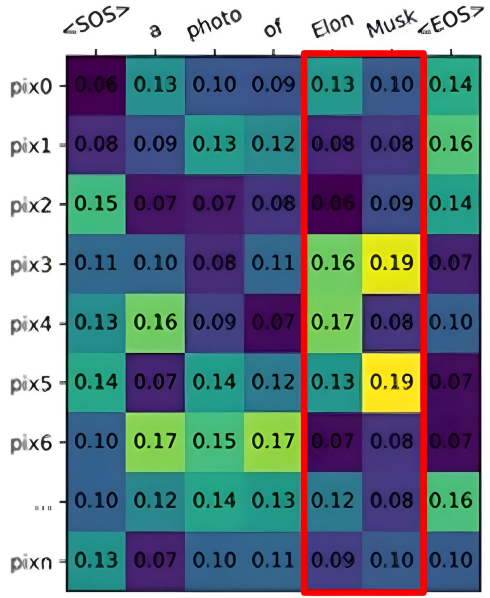
$$h = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

Attention Re-steering Loss



$$\mathcal{L}_{Attn} = \sum_{a \in A[:,i:j]} ||a||^2$$

Attention Re-steering Loss



$$\mathcal{L}_{Attn} = \sum_{a \in A[:,i:j]} ||a||^2$$

Visual Denoising Loss

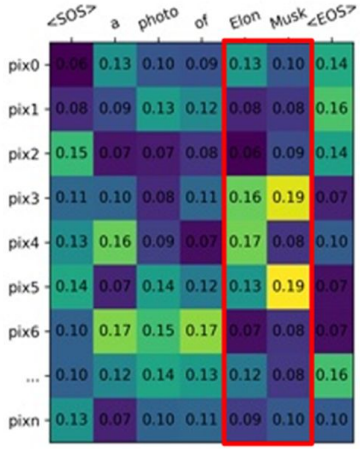


$\hat{x}_0$

Visual Denoising Loss



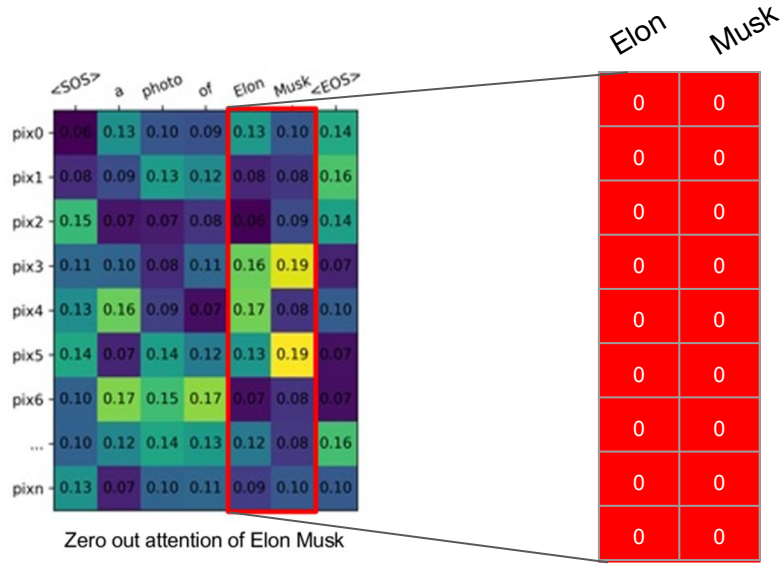
$\hat{x}_0$



Visual Denoising Loss



$\hat{x}_0$

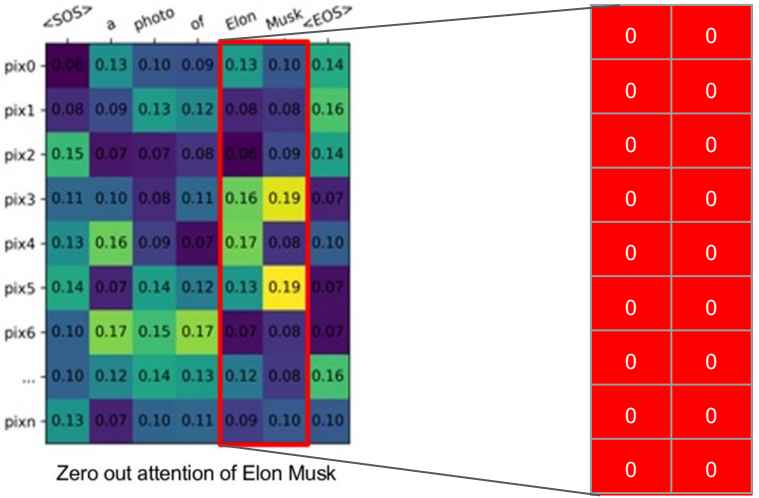


Visual Denoising Loss



$\hat{x}_0$

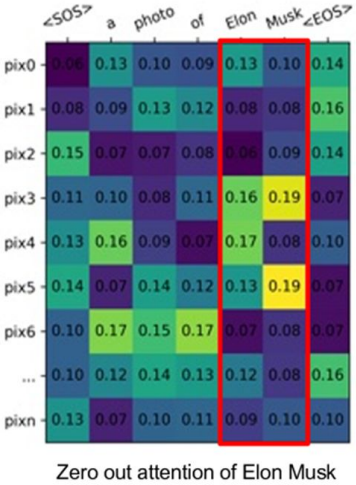
$\tilde{x}_0$



Visual Denoising Loss



$\hat{x}_0$



Zero out attention of Elon Musk

$\tilde{x}_0$

$$\mathcal{L}_{vis} = \mathrm{MSE}(\hat{x}_0, \tilde{x}_0)$$

Two Losses

Attention Re-steering Loss

Two Losses

Attention Re-steering Loss

OR

Visual Denoising Loss

Two Losses

Attention Re-steering Loss

OR

Visual Denoising Loss

OR

Both

Textual Inversion

For concepts which can’t be described in text



Input samples

Textual Inversion



Input samples $\xrightarrow{invert}$ "S_*"

Assigning a special token

Textual Inversion



Input samples $\xrightarrow{invert}$ "S_*"

Assigning a special token

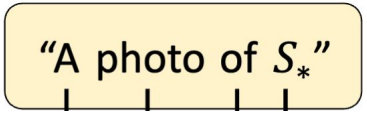


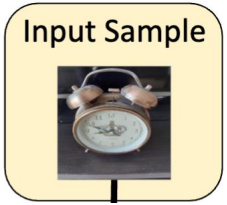
"An oil painting of S_*"



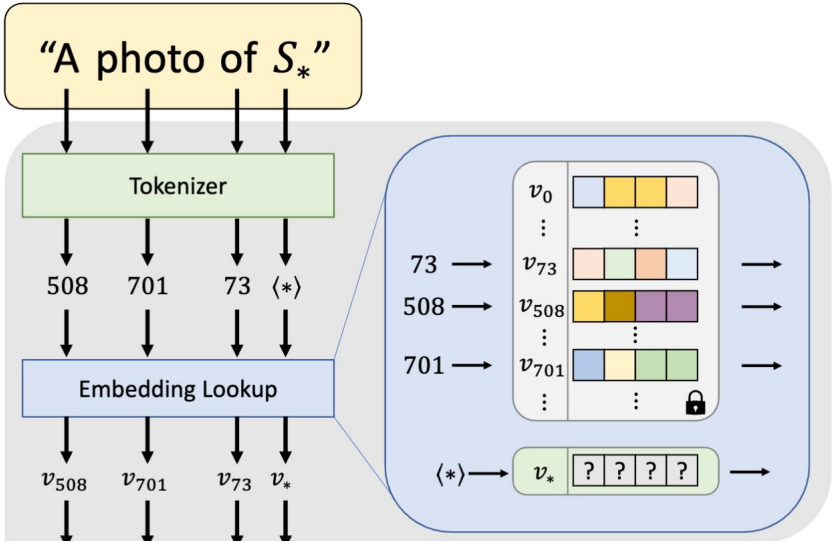
"App icon of S_*"

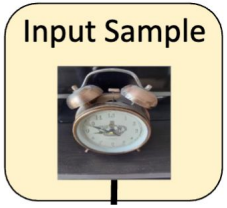
Textual Inversion Process



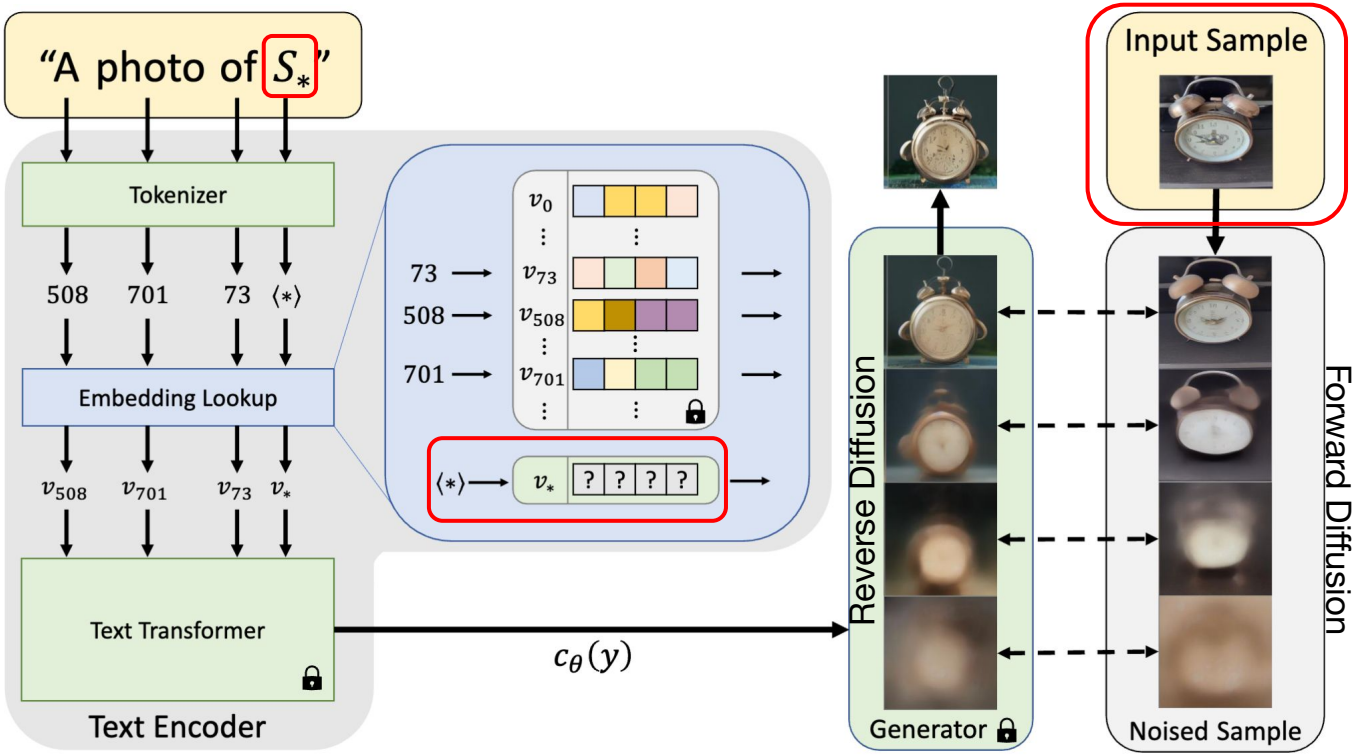


Textual Inversion





Textual Inversion

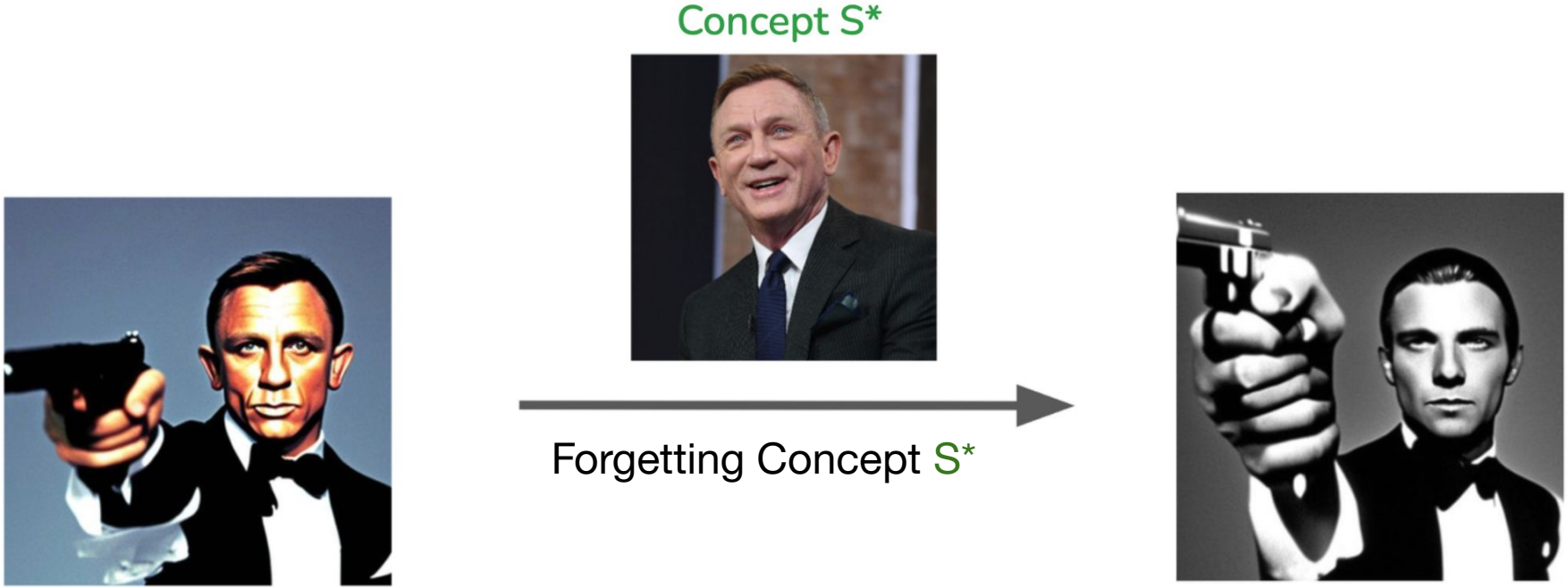


Concept Inversion

Concept S*



Concept Inversion



Evaluation

CLIP Score

Evaluation

CLIP Score

Memorization Score

Evaluation - CLIP Score

cosine similarity between image and text embeddings



⇔ Elon Musk

CLIP Score
(0.308)



⇔ Elon Musk

CLIP Score
(0.285)

Evaluation - CLIP Score

cosine similarity between two embeddings

Concept
Elon Musk
Mickey Mouse
Zebra
Google
Apple
Horse
Van Gogh

Evaluation - CLIP Score

cosine similarity between two embeddings

Concept	Init. CLIP
Elon Musk	0.308
Mickey Mouse	0.304
Zebra	0.312
Google	0.216
Apple	0.267
Horse	0.275
Van Gogh	0.274

Evaluation - CLIP Score

cosine similarity between two embeddings

Concept	Init. CLIP	Frgt. CLIP ↓
Elon Musk	0.308	0.285
Mickey Mouse	0.304	0.269
Zebra	0.312	0.310
Google	0.216	0.209
Apple	0.267	0.258
Horse	0.275	0.266
Van Gogh	0.274	0.233

Evaluation - Memorization Score

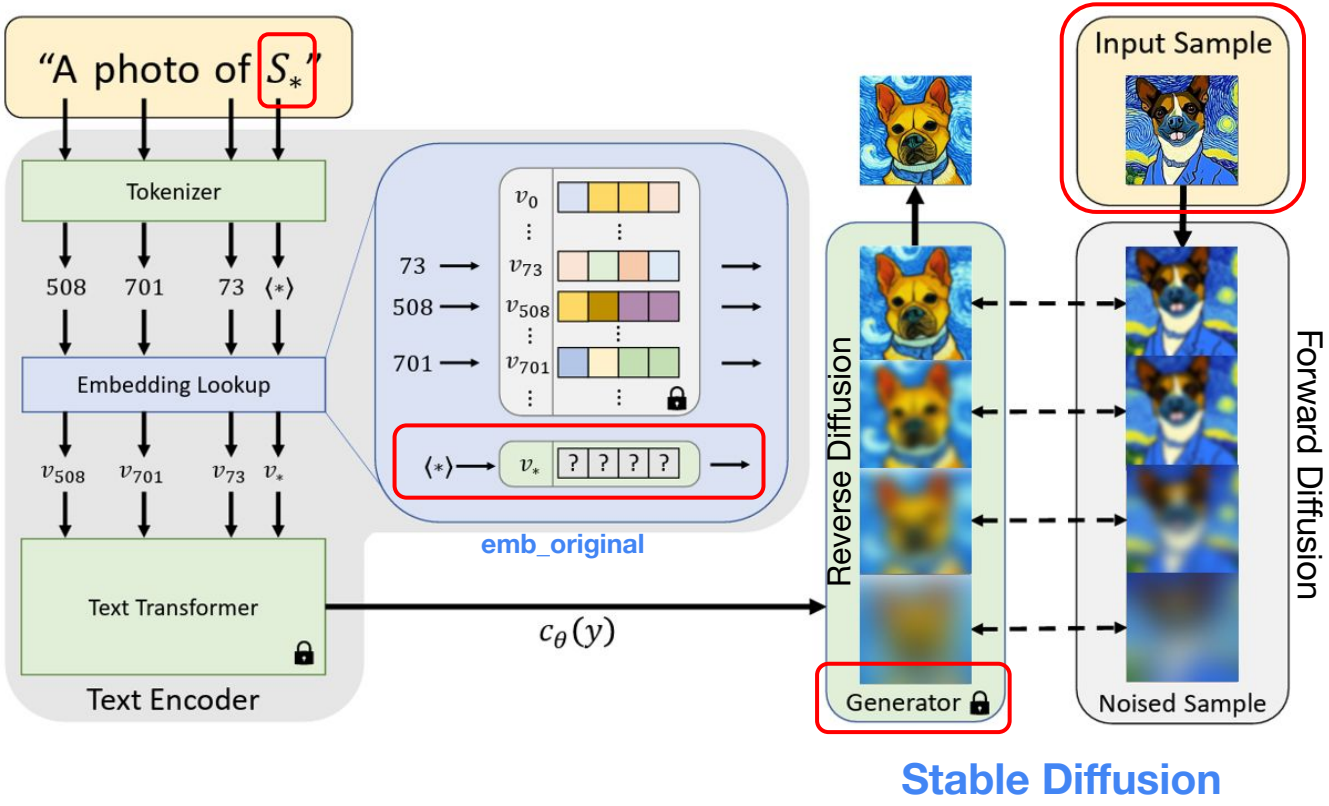
$$\text{Init MScore} = \text{Cos}\left(\text{Emb_Anchor}, \text{Emb_Original} \right)$$

Evaluation - Memorization Score

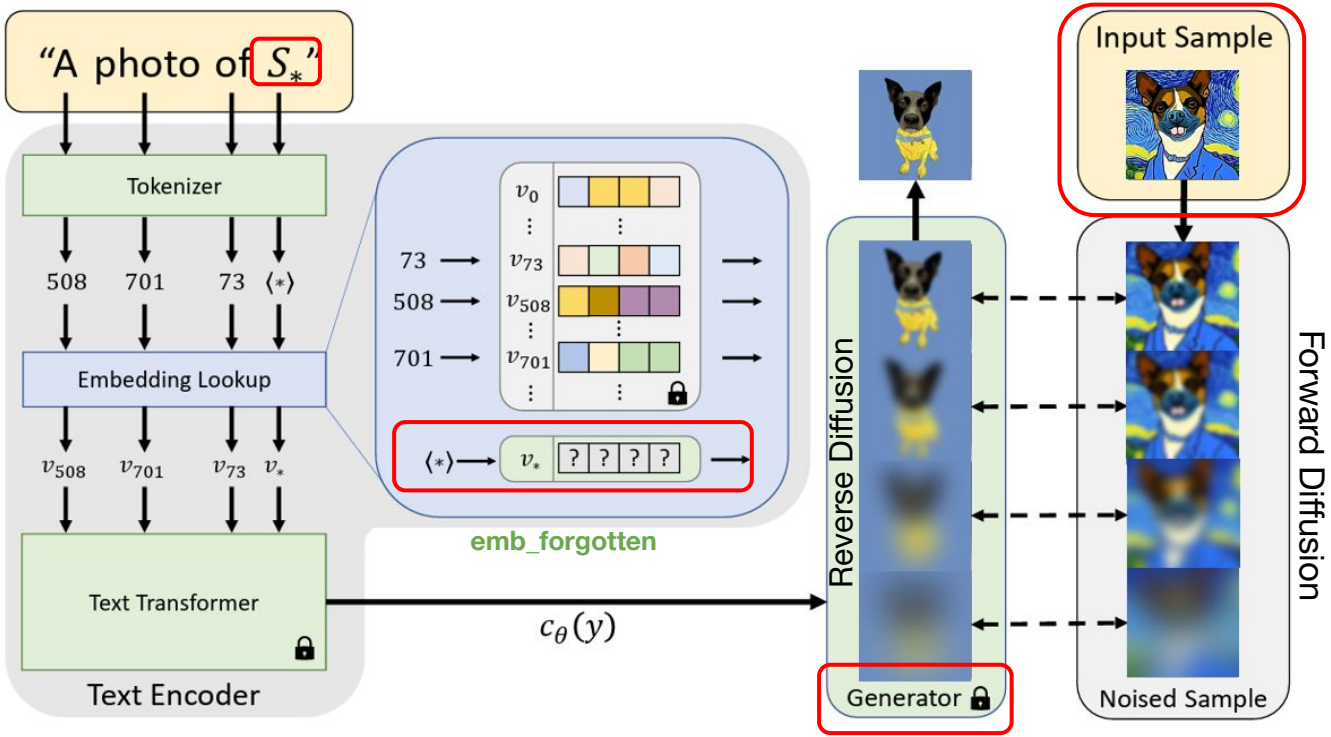
$$\text{Init MScore} = \text{Cos}\left(\text{Emb_Anchor}, \text{Emb_Original}\right)$$

$$\text{Frgt MScore} = \text{Cos}\left(\text{Emb_Anchor}, \text{Emb_Forgotten}\right)$$

emb_original using textual inversion



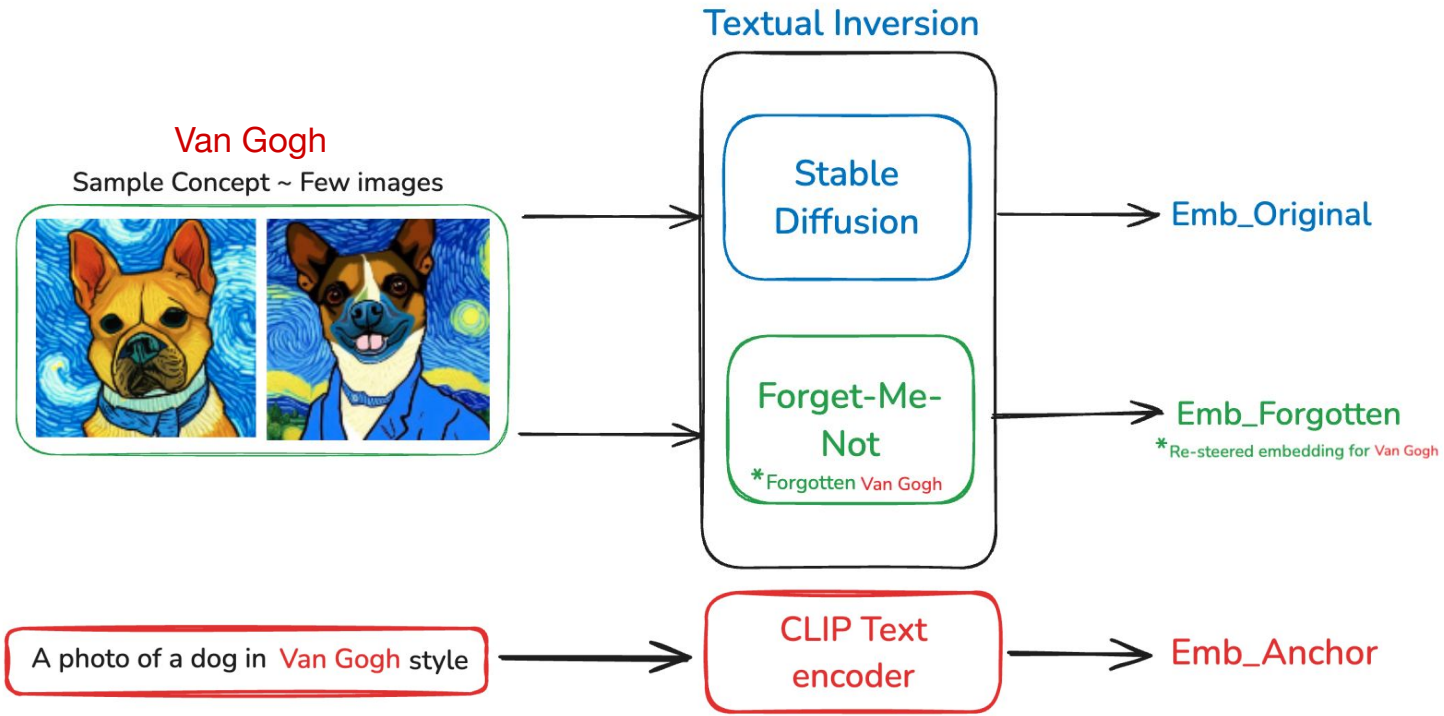
emb_forgotten using textual inversion



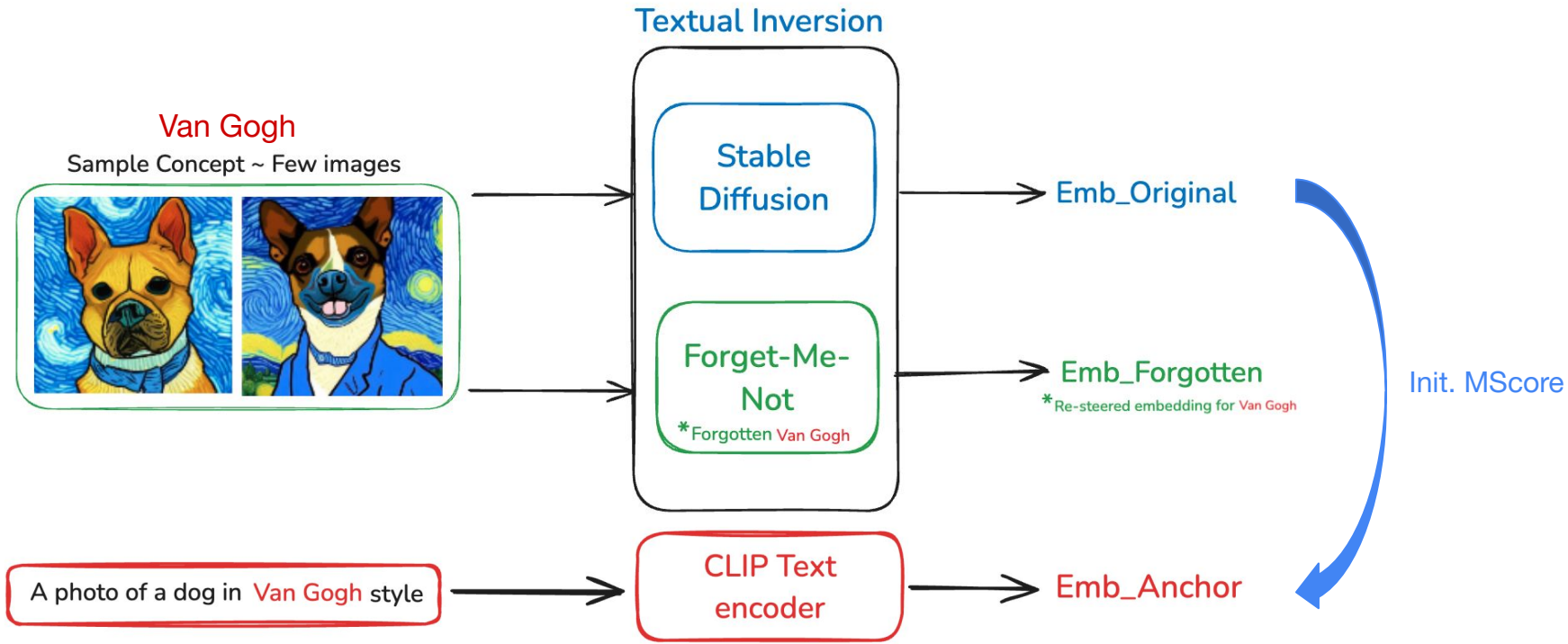
Forget Me Not Model

(forgotten Van Gogh)

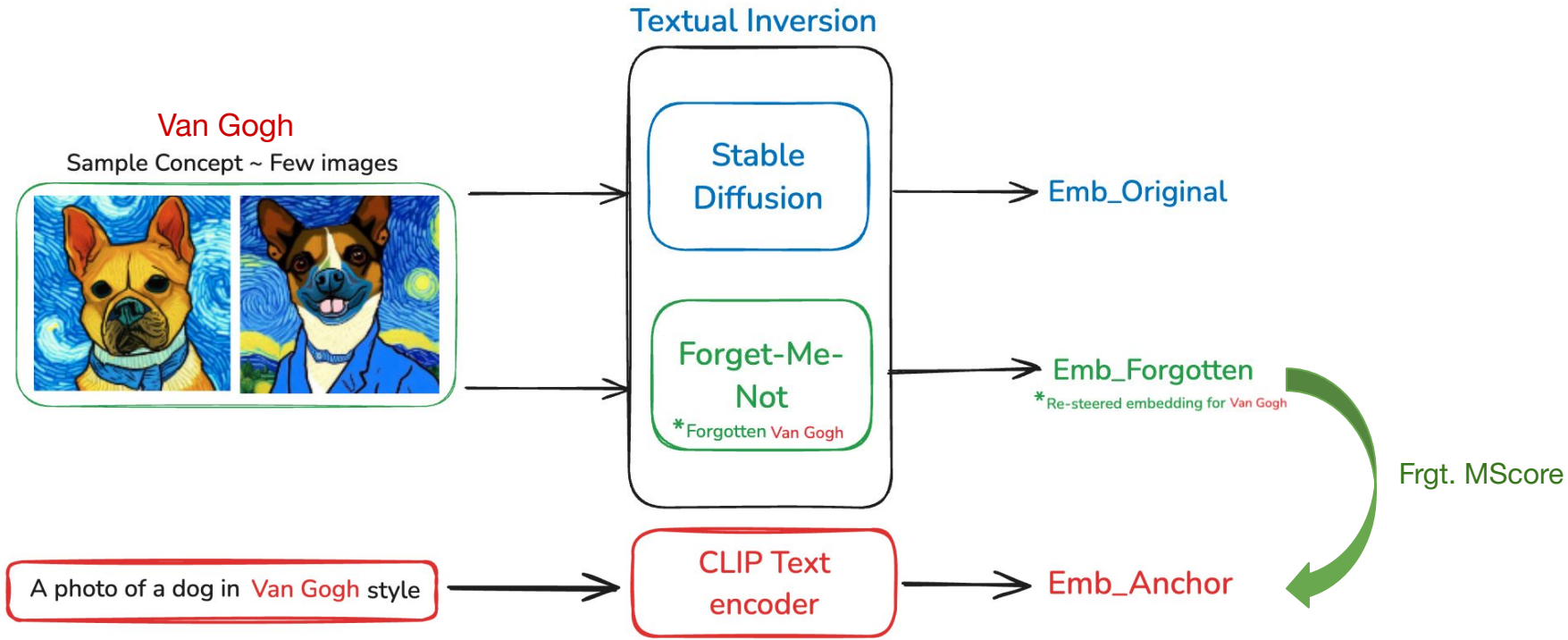
Evaluation - Memorization Score



Evaluation - Memorization Score



Evaluation - Memorization Score



Evaluation - Memorization Score

Concept
Elon Musk
Mickey Mouse
Zebra
Google
Apple
Horse
Van Gogh

Evaluation - Memorization Score

Concept	Init. MScore
Elon Musk	0.943
Mickey Mouse	0.948
Zebra	0.972
Google	0.940
Apple	0.696
Horse	0.877
Van Gogh	0.916

Evaluation - Memorization Score

Concept	Init. MScore	Frgt. MScore ↓
Elon Musk	0.943	0.848
Mickey Mouse	0.948	0.836
Zebra	0.972	0.899
Google	0.940	0.811
Apple	0.696	0.493
Horse	0.877	0.808
Van Gogh	0.916	0.684

Qualitative results - Multi-concept model

Elon Musk



Qualitative results - Multi-concept model

Elon Musk





Qualitative results - Multi-concept model





Qualitative results - Multi-concept model

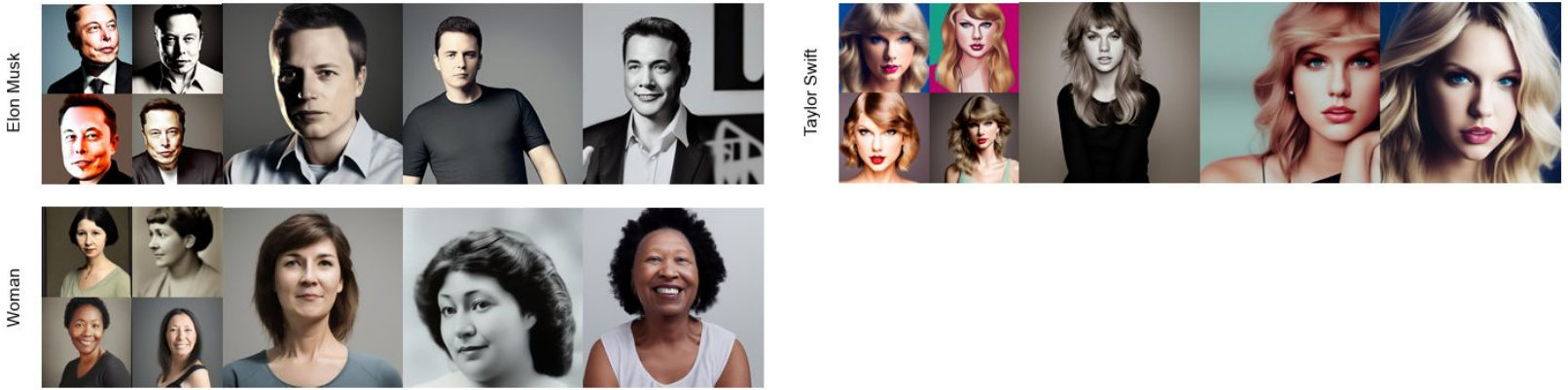
Elon Musk



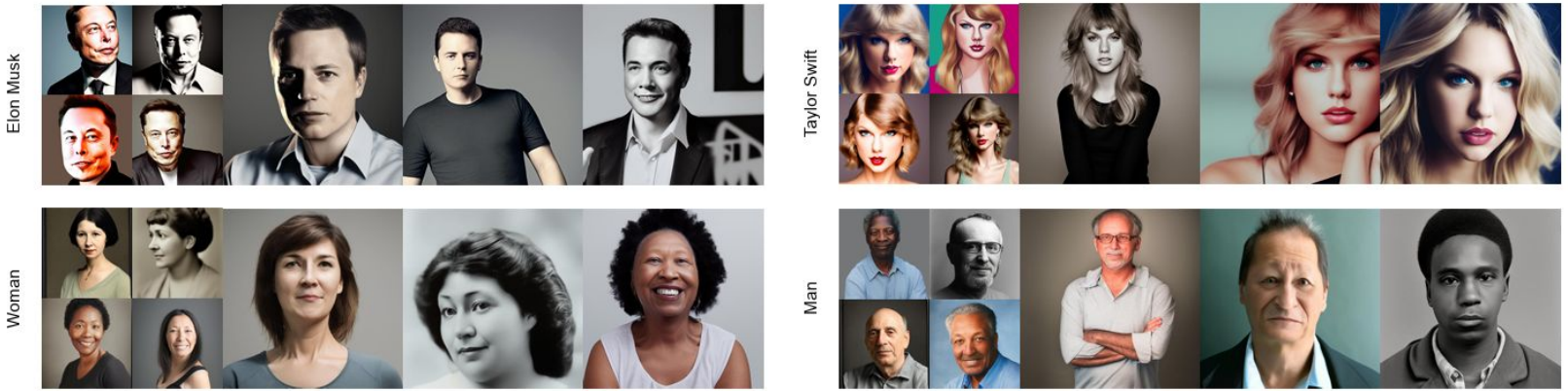
Taylor Swift



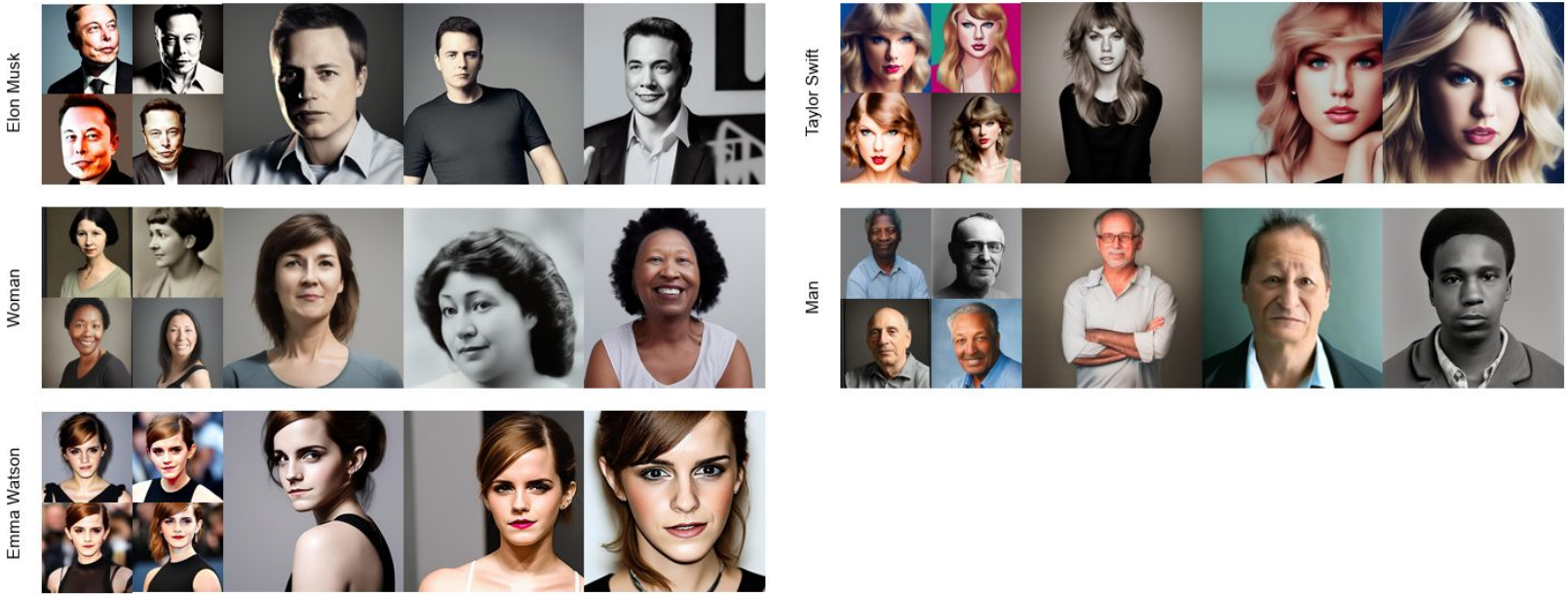
Qualitative results - Multi-concept model



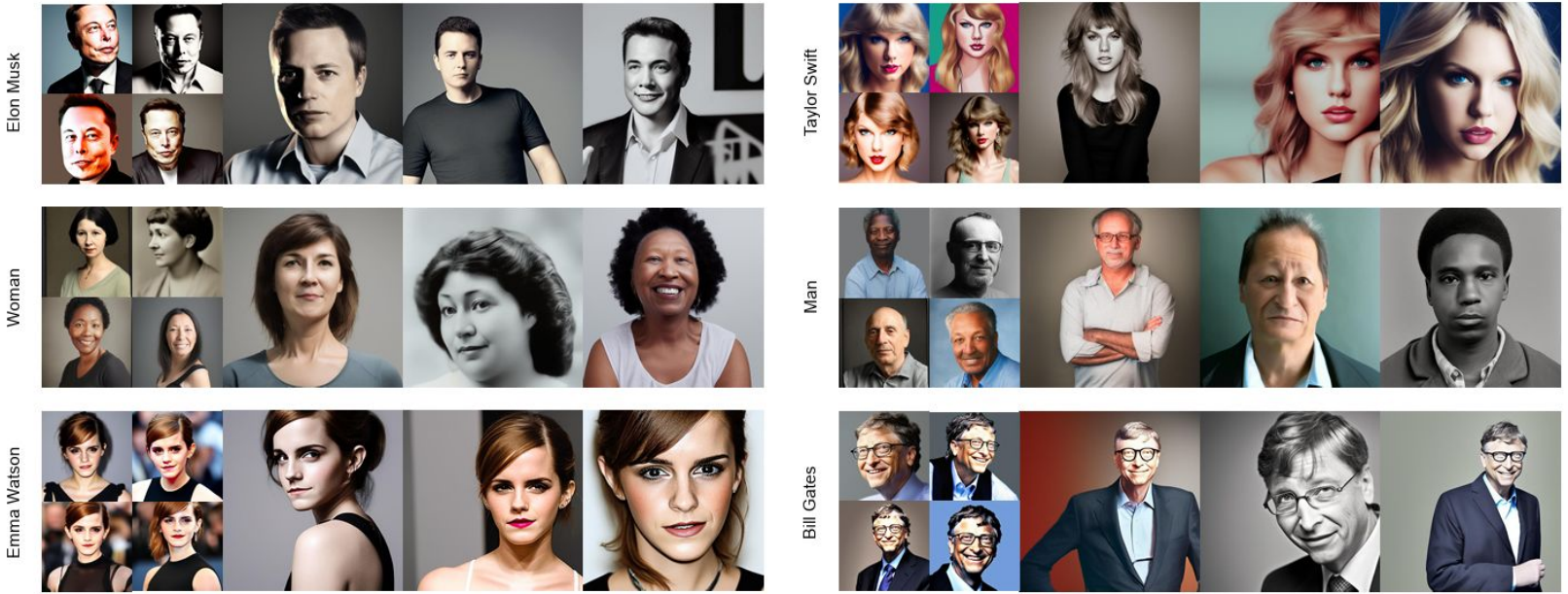
Qualitative results - Multi-concept model



Qualitative results - Multi-concept model



Qualitative results - Multi-concept model



Qualitative results - Style



Qualitative results - Style



Qualitative results - Concept Correction

Dominant Concept



Qualitative results - Concept Correction



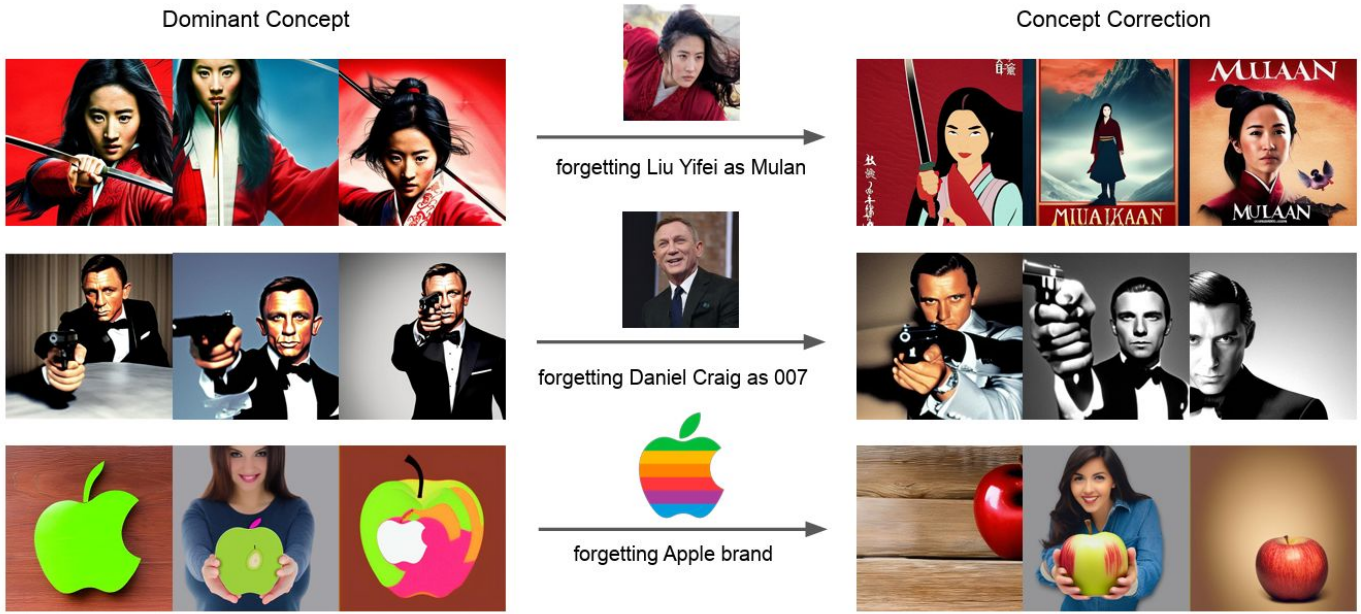
Qualitative results - Concept Correction



Qualitative results - Concept Correction



Qualitative results - Concept Correction



Qualitative results - Concept Inversion



Qualitative results - Concept Inversion



Qualitative results - Concept Inversion



Concept Inversion is not always ideal

Orignal: A photo of airplane



Concept Inversion is not always ideal

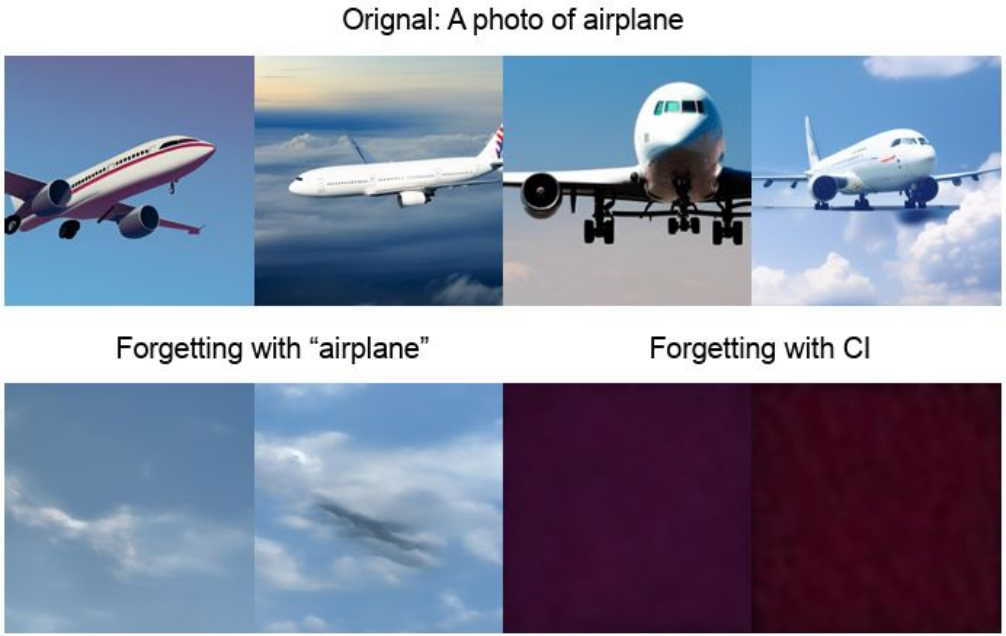
Orignal: A photo of airplane



Forgetting with “airplane”



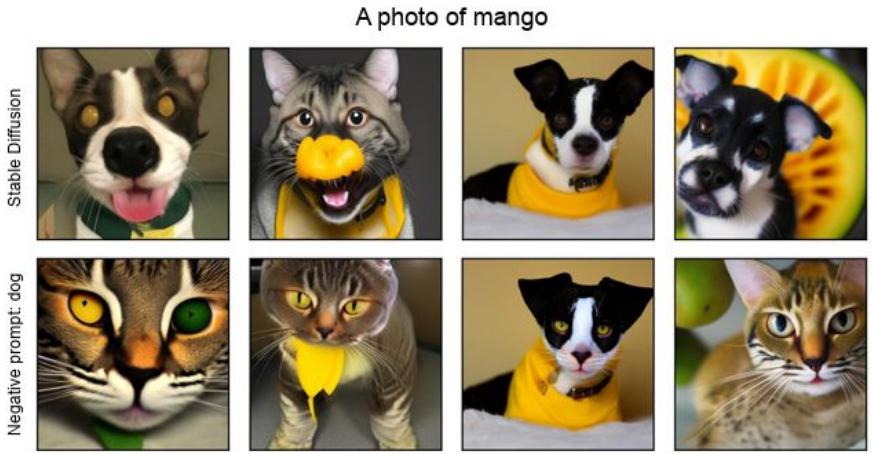
Concept Inversion is not always ideal



Concept Correction



Concept Correction



Concept Correction

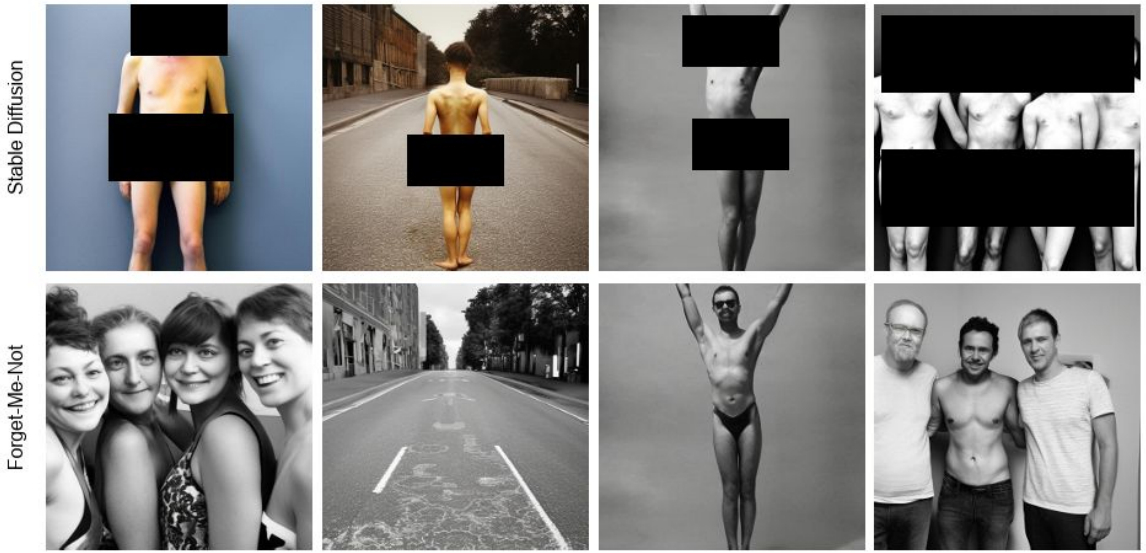


Qualitative results - NSFW Images

Stable Diffusion



Qualitative results - NSFW Images



Comparison - forgetting methods

Task - Trying to forget **french horn**

“a photo of a violin”, “a photo of a french horn”

Comparison - forgetting methods

Task - Trying to forget **french horn**

“a photo of a violin”, “a photo of a french horn”



SD

Comparison - forgetting methods

Task - Trying to forget **french horn**

“a photo of a violin”, “a photo of a french horn”



SD



ESD

Comparison - forgetting methods

Task - Trying to forget **french horn**

“a photo of a violin”, “a photo of a french horn”



SD



ESD



ACTD

Comparison - forgetting methods

Task - Trying to forget **french horn**

“a photo of a violin”, “a photo of a french horn”



SD



ESD



ACTD

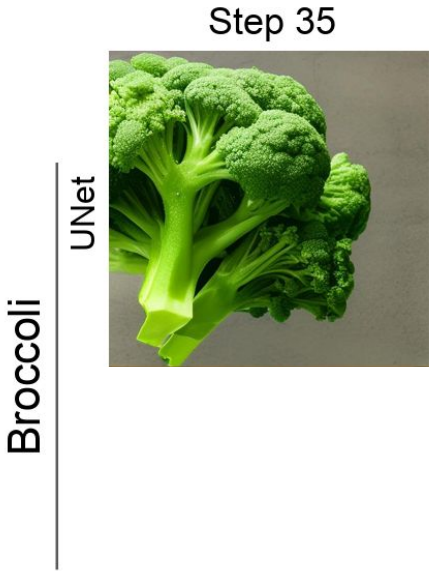


Proposed

Ablation - all weights vs cross-attention weights



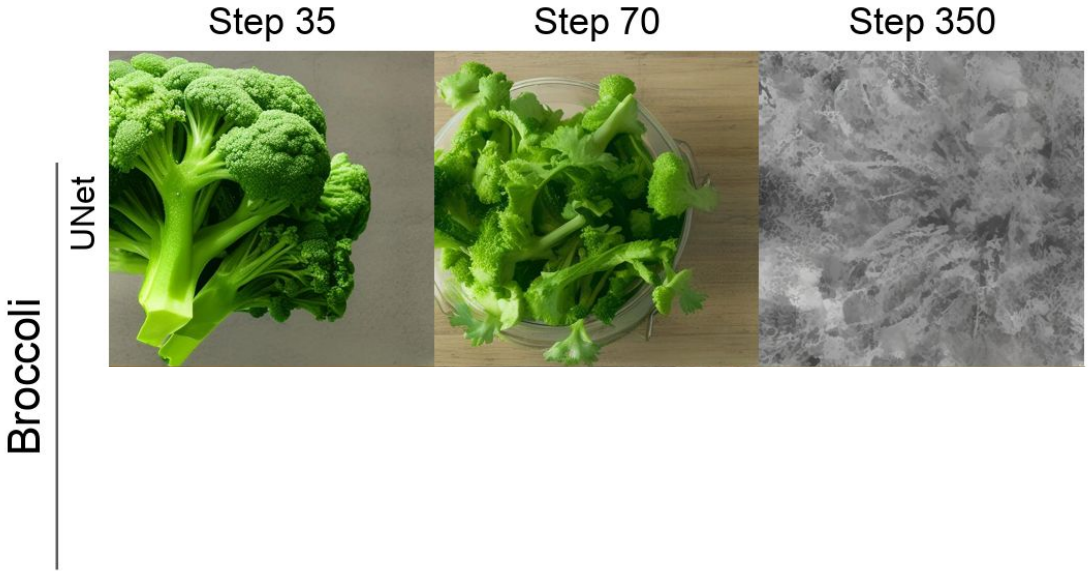
Ablation - all weights vs cross-attention weights



Ablation - all weights vs cross-attention weights

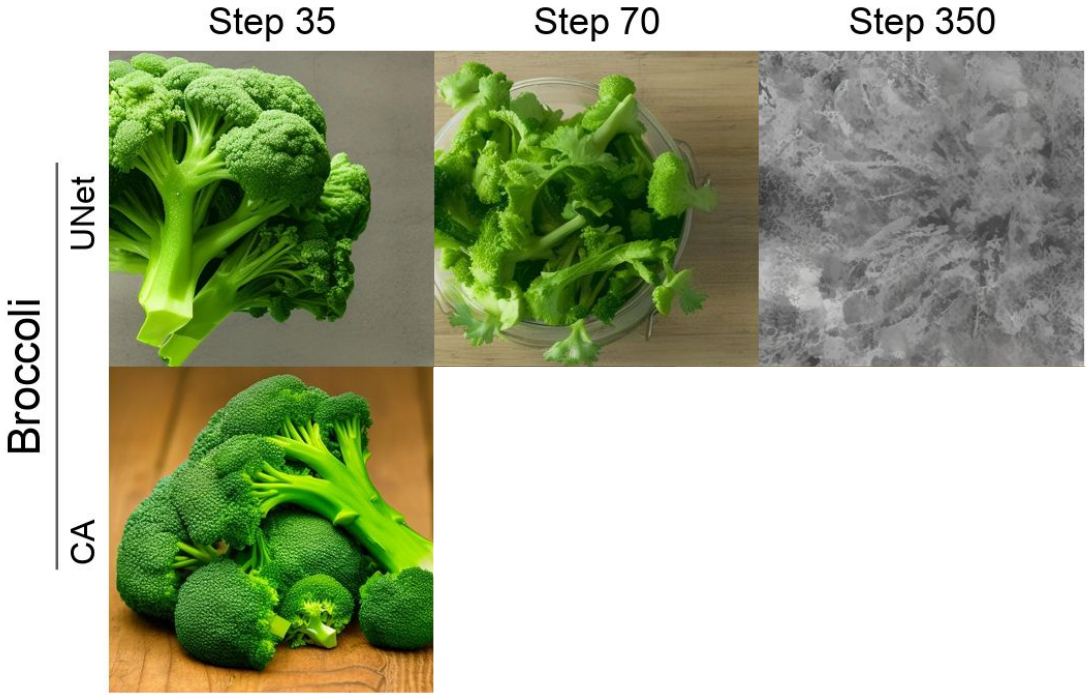


Ablation - all weights vs cross-attention weights

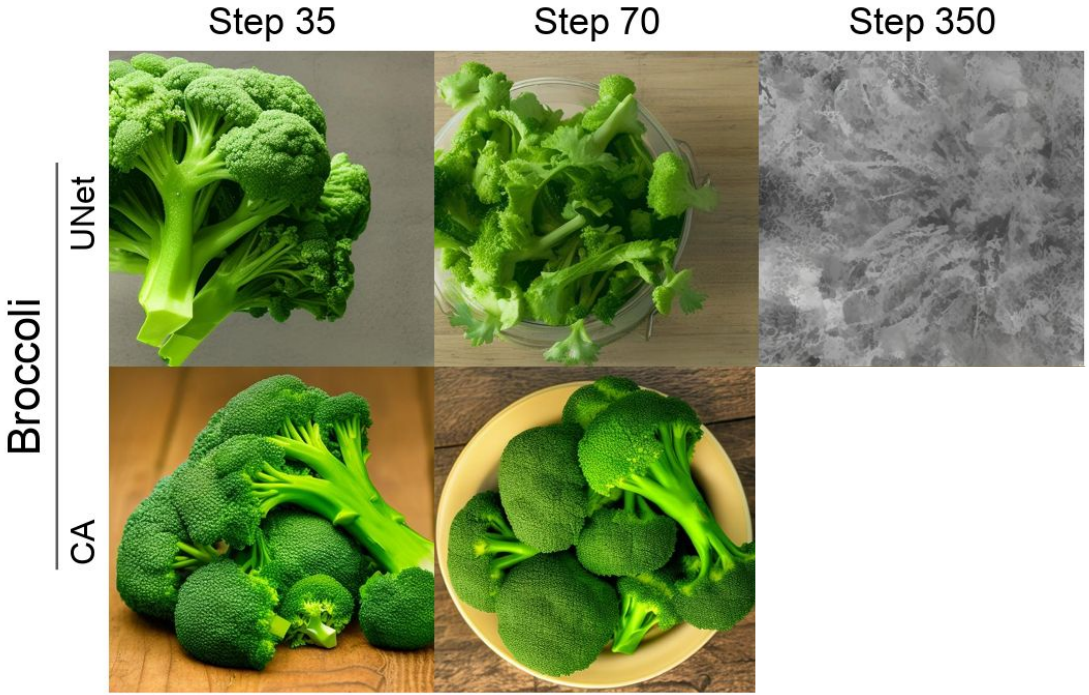


Model collapse

Ablation - all weights vs cross-attention weights



Ablation - all weights vs cross-attention weights



Ablation - all weights vs cross-attention weights



Ablation - all weights vs cross-attention weights



Testing model at 350 steps

Ablation - all weights vs cross-attention weights



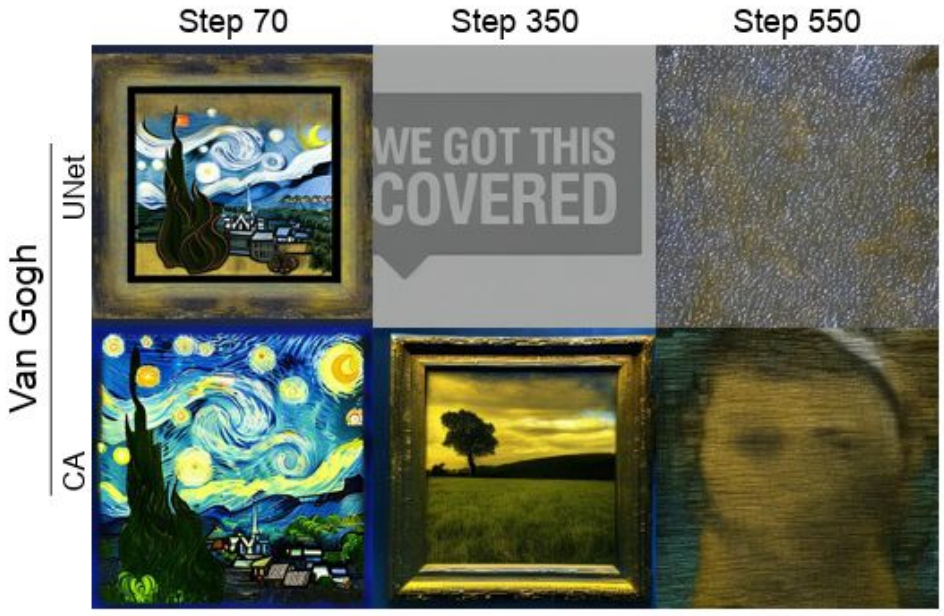
A photo of Elon musk

Ablation - all weights vs cross-attention weights

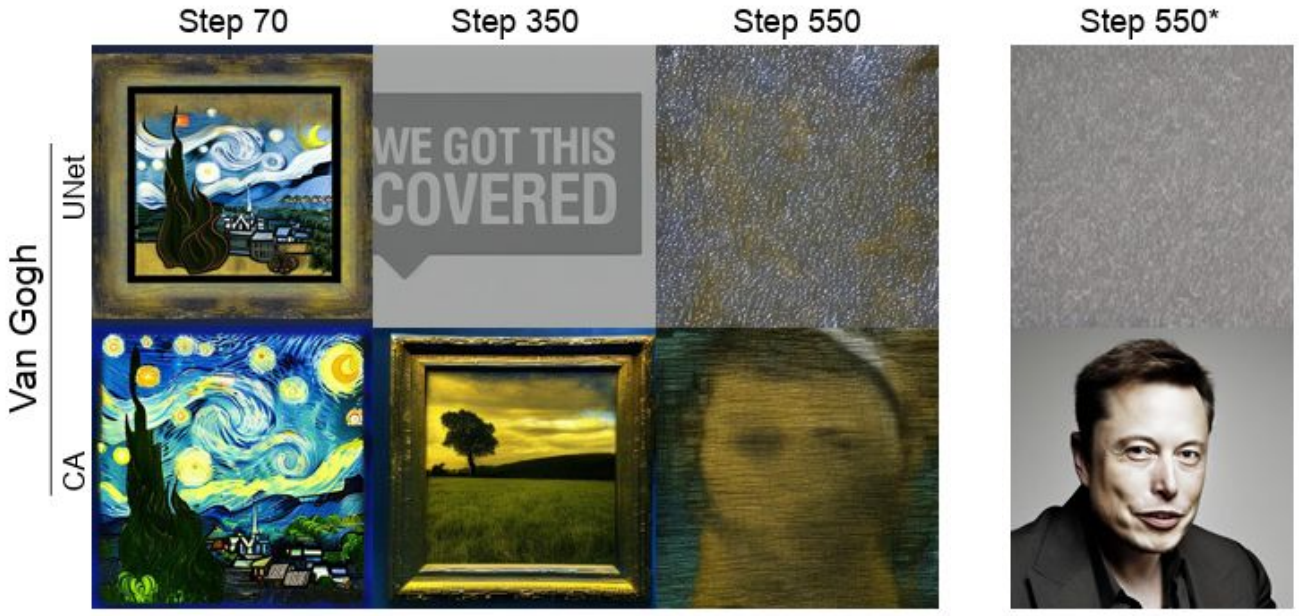


A photo of Elon musk

Ablation - all weights vs cross-attention weights



Ablation - all weights vs cross-attention weights



Unrelated Concept

(elon musk)

Limitations

- Resteering loss or Visual Denoising loss or both

Limitations

- Resteering loss or Visual Denoising loss or both
  - The losses are very similar and without ablations

Limitations

- Resteering loss or Visual Denoising loss or both
  - The losses are very similar and without ablations

- Concept Inversion is introduced without ablation on the larger dataset

Limitations

- Resteering loss or Visual Denoising loss or both
  - The losses are very similar and without ablations
- Concept Inversion is introduced without ablation on the larger dataset
- The number of concepts in the result table are very little

Conclusion

- Overall its a light-weight solution for concept forgetting.

Conclusion

- Overall its a light-weight solution for concept forgetting.

- The ablation for augmented query text is missing (e.g. after forgetting “elon musk” what if the prompt is given like “the owner of tesla/spacex company”?

Conclusion and Limitations

- Overall its a light-weight solution for concept forgetting.
- Struggle to forget abstract concepts.
- Restreering loss or Visual Denoising loss or both
  - The losses are very similar and without ablations
- Concept Inversion is introduced without ablation on the larger dataset
- The number of concepts in the result table are very little
- Figure 3 could’ve been better, the matmul and attn operations are not properly shown.
- The ablation for augmented query text is missing (e.g. after forgetting “elon musk” what if the prompt is given like “the owner of tesla/spacex company”?
- Introduced Mscore is not properly reflected in results table.

Thank You!