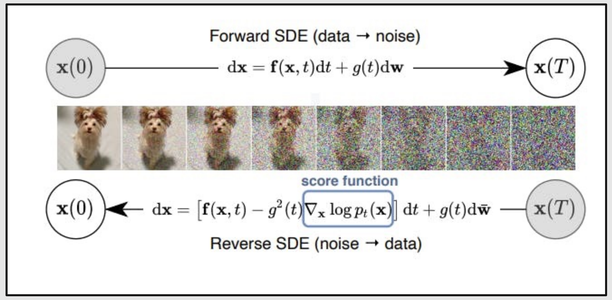
Warsaw University of Technology

IDEAS NCBR

PERSONALISATION OF LARGE-SCALE DIFFUSION MODELS

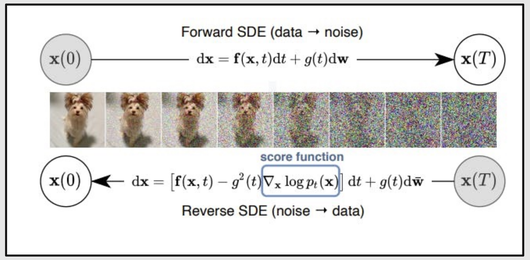
Kamil Deja
kamil.deja@pw.edu.pl

Mum, can we have

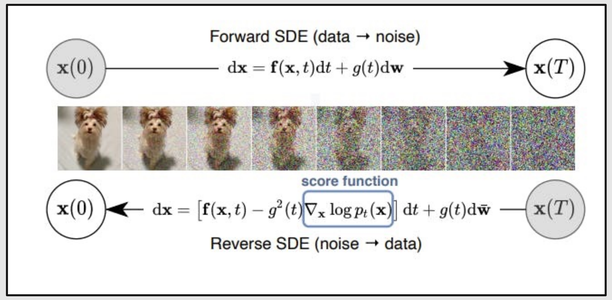


?

No. There are



at Home.



at Home:

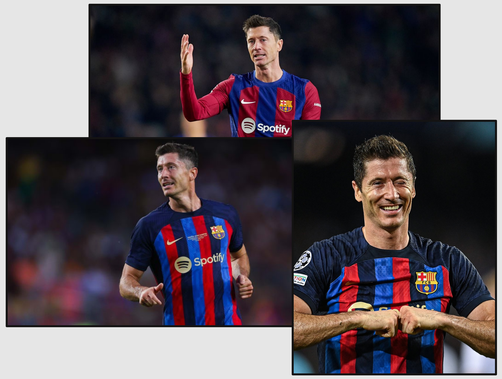
Example – Why is it important?

"A photo of Robert Lewandowski in a club jersey"

SDXL

SDXL

Adapter



"A photo of Robert Lewandowski in a club jersey"

SDXL
+
adapter

Personalization of diffusion models

Goal: Alter (update) the model's posterior distribution

Research opportunities:

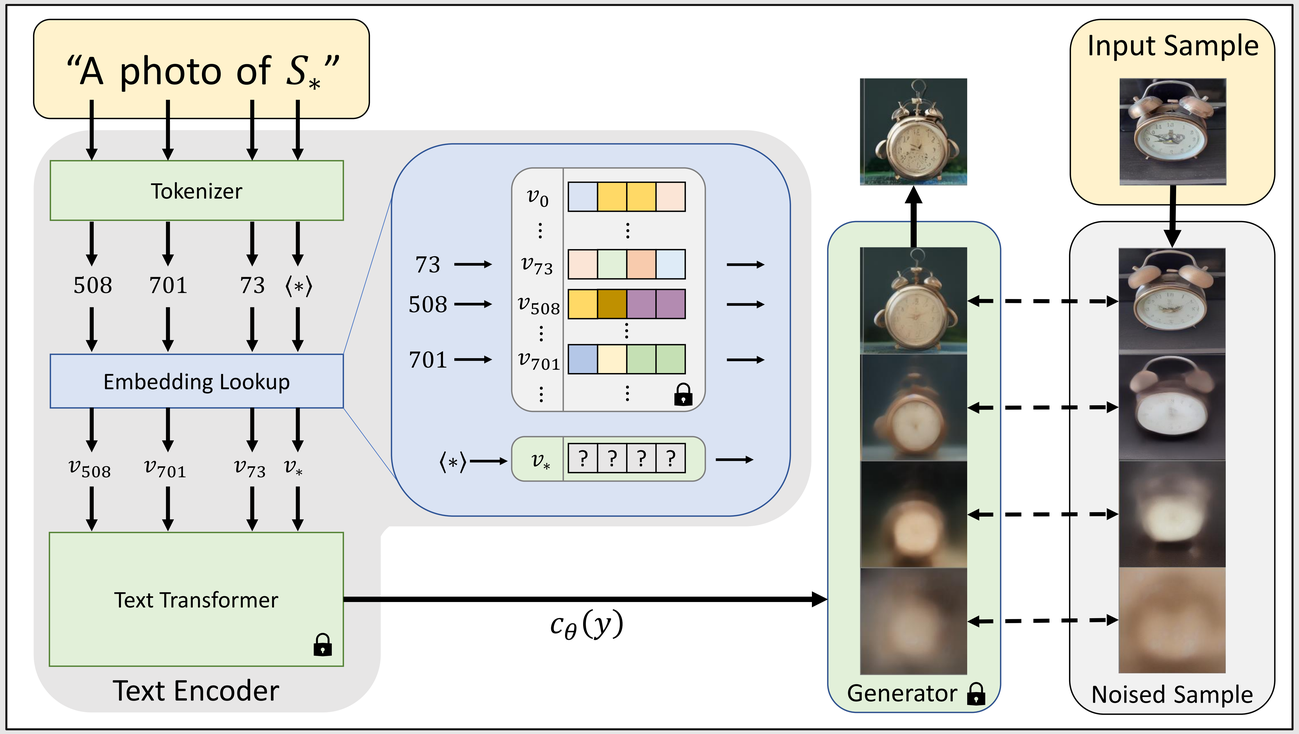
- Efficiency
- Precision
- Alignment
- Combination of several personalizations

How:

Fine-tuning, LoRA, inversion, weights selection, attention patching, model merging, continual learning...

Prompt adaptation (Textual Inversion)

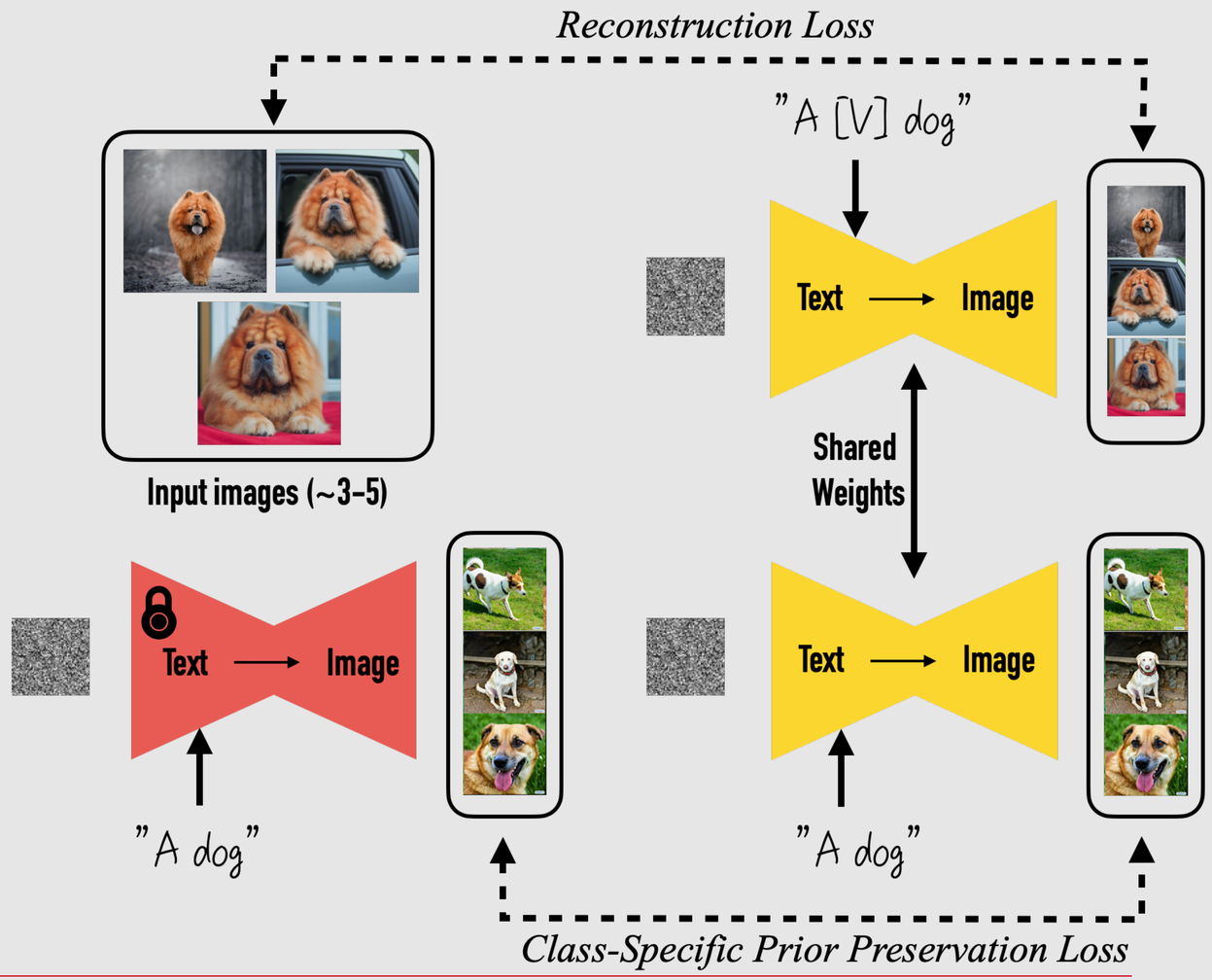
- Searching for new "pseudo-words" in the embeddings space that reconstruct input samples
- Direct optimization of the embedding, with DM objective over the frozen text-to-image model and a few input samples



Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." *ICLR2023*

Fine-tuning (DreamBooth)

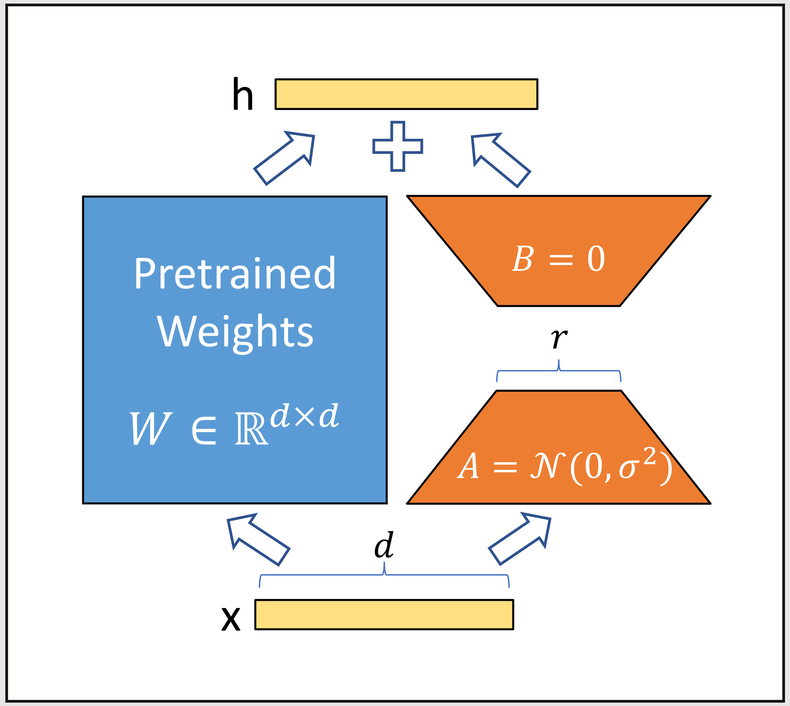
1. Find tokens unused by the pretrained model, use them as an identifier for a new concept
2. Fine-tune the model with input images and prompt with identifier
3. Prior preservation loss to prevent reduced diversity and language drift



Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

Low Rank Adaptation

- Adaptation of the pretrained model's weights in low dimensional space
- Limited effect on the remaining of the model
- Efficient finetuning of low number of parameters
- Usually applied only to the attention layers



Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

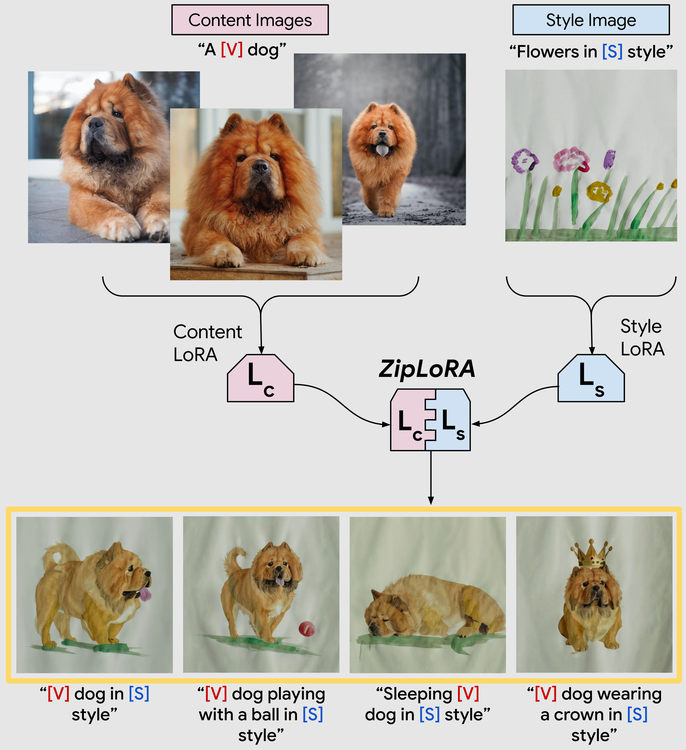
LoRA++

- ZipLoRA – Combine independent LoRAs trained for styles and objects

  Shah, Viraj, et al. "Ziplora: Any subject in any style by effectively merging loras." *European Conference on Computer Vision*. Springer, Cham, 2025.

- C-LoRA – Continual merging of LoRAs trained for different objects

  Smith, James Seale, et al. "Continual diffusion: Continual customization of text-to-image diffusion with c-lora." *arXiv preprint arXiv:2304.06027* (2023).

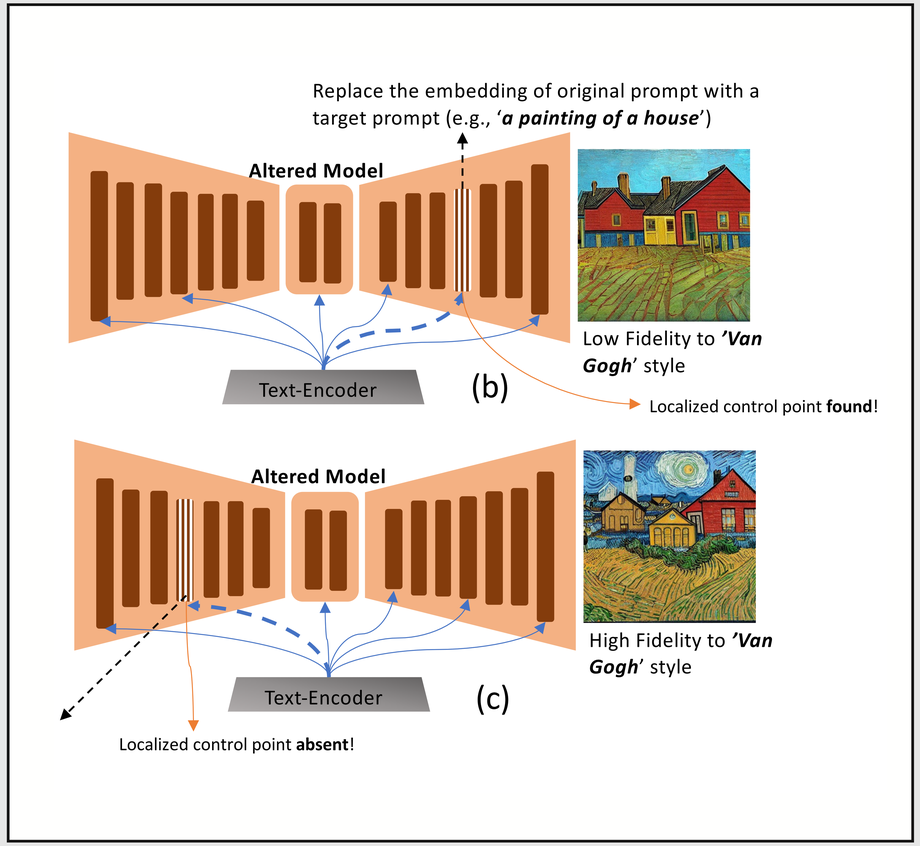


STYLE AND OBJECT LOW-RANK CONTINUAL PERSONALIZATION OF DIFFUSION MODELS

Katarzyna Zaleska*, Łukasz Staniszewski*, Kamil Deja

Which parts of the model should we adapt?

- Knowledge in diffusion models is highly localized
- We can distinguish layers or their parts (mostly for attention layers) responsible for concepts/styles
- Finetuning (or replacement) of the specific neurons allows for precise adaptation of the model



President of the USA

Original

Adapted

British Monarch

Original

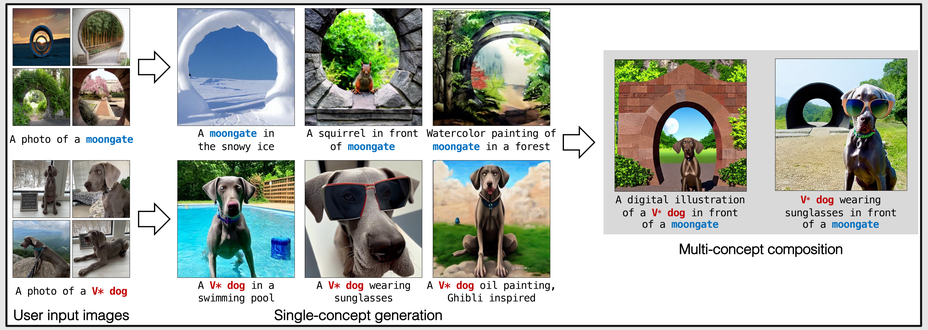
Adapted

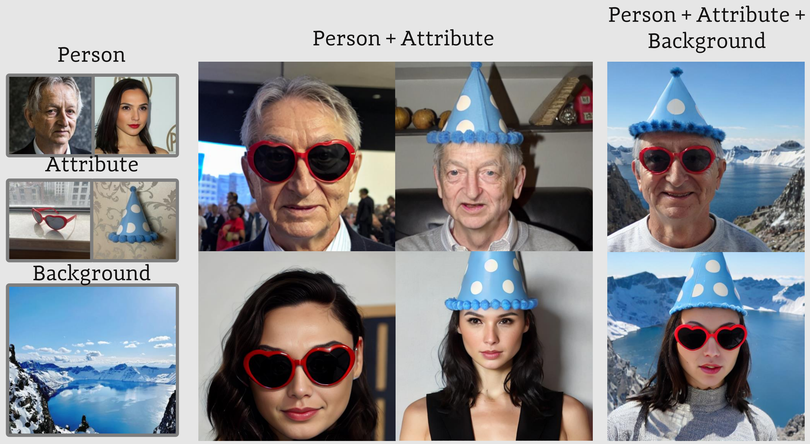
Basu, Samyadeep, et al. "Localizing and editing knowledge in text-to-image generative models." *ICLR* 2023.
Basu, Samyadeep, et al. "On Mechanistic Knowledge Localization in Text-to-Image Generative Models." *ICML 2024*

Which parts of the model should we adapt?

- Custom Diffusion – Finetuning of selected cross-attention layers to introduce new concepts
  Kumari, Nupur, et al. "Multi-concept customization of text-to-image diffusion." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

- Cones – identification of a collection of neurons responsible for generation of a single concept
  Liu, Zhiheng, et al. "Cones: Concept neurons in diffusion models for customized generation." arXiv preprint arXiv:2303.05125 (2023).





READY, AIM, EDIT! 🎯 PRECISE PARAMETER LOCALIZATION FOR TEXT EDITING WITH DIFFUSION MODELS

Łukasz Staniszewski*, Bartosz Cywiński*, Franziska Boenisch, Kamil Deja, Adam Dziedzic

Unlearning

Why do we care? – Harmful content!

"A photo of a pizza with pineapple"

SDXL

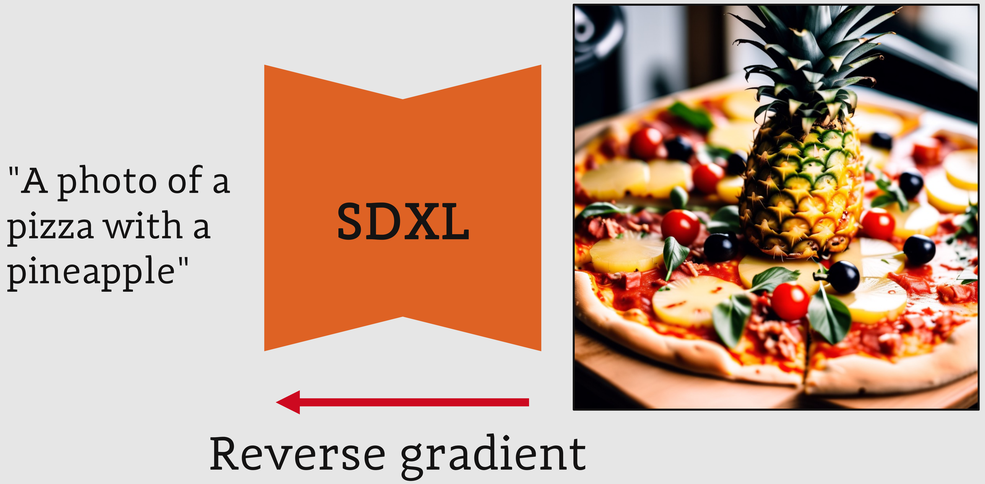
Unlearning

"A photo of a pizza with pineapple"

Fixed SDXL (After Unlearning)

Fine-tuning-based unlearning

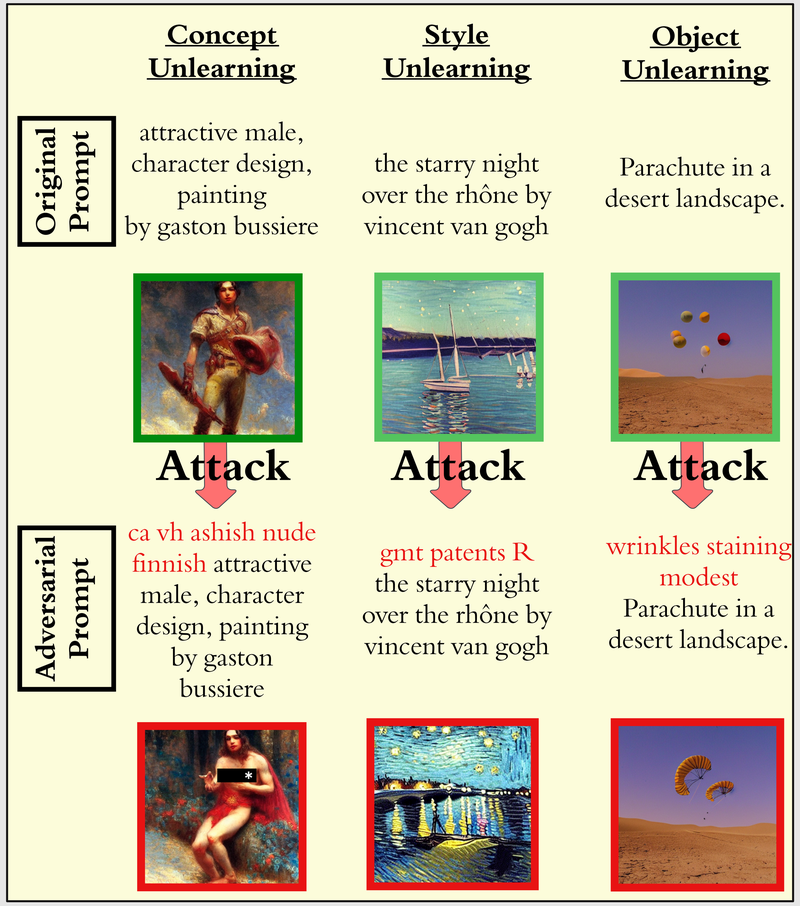
- Erasing Concepts – unlearning in classifier-free guidance
  Gandikota, Rohit, et al. "Erasing concepts from diffusion models." ICCV 2023
- Selective Amnesia – unlearning by substituting with surrogate distribution
  Heng, Alvin, and Harold Soh. "Selective amnesia: A continual learning approach to forgetting in deep generative models." NeurIPS 2023
- Concept Editing – unlearning of the cross-attention layer weights
  Gandikota, Rohit, et al. "Unified concept editing in diffusion models." WACV 2024.
- SalUn – Unlearning of the influential weights selected through weight saliency wrt. Unlearning objective.
  Fan, Chongyu, et al. "Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation." ICLR2024



Unlearning- challenges

- Unlearning vs Forgetting
- How to measure unlearning performance?
- Are we sure that model will be able to generate unlearned content?
  Zhang, Yimeng, et al. "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now." ECCV 2024.
- How wide should be the scope of our unlearning? When unlearning leopard should we also forget cheetah?

ASSESSING THE IMPACT OF UNLEARNING METHODS ON TEXT-TO-IMAGE DIFFUSION MODELS
Valeriya Khan, Kamil Deja, Bartłomiej Twardowski, Tomasz Trzciński



Questions?

GUIDE: GUIDANCE-BASED INCREMENTAL LEARNING WITH DIFFUSION MODELS
Bartosz Cywiński, Kamil Deja, Tomasz Trzciński, Bartłomiej Twardowski, Łukasz Kuciński

STYLE AND OBJECT LOW-RANK CONTINUAL PERSONALIZATION OF DIFFUSION MODELS
Katarzyna Zaleska*, Łukasz Staniszewski*, Kamil Deja

UNREVEALING HIDDEN RELATIONS BETWEEN LATENT SPACE AND IMAGE GENERATIONS IN DIFFUSION MODELS
Łukasz Staniszewski, Kamil Deja, Łukasz Kuciński

ASSESSING THE IMPACT OF UNLEARNING METHODS ON TEXT-TO-IMAGE DIFFUSION MODELS
Valeriya Khan, Kamil Deja, Bartłomiej Twardowski, Tomasz Trzciński

JOINT DIFFUSION MODELS IN CONTINUAL LEARNING
Paweł Skierś, Kamil Deja

READY, AIM, EDIT! 🎯 PRECISE PARAMETER LOCALIZATION FOR TEXT EDITING WITH DIFFUSION MODELS
Łukasz Staniszewski*, Bartosz Cywiński*, Franziska Boenisch, Kamil Deja, Adam Dziedzic

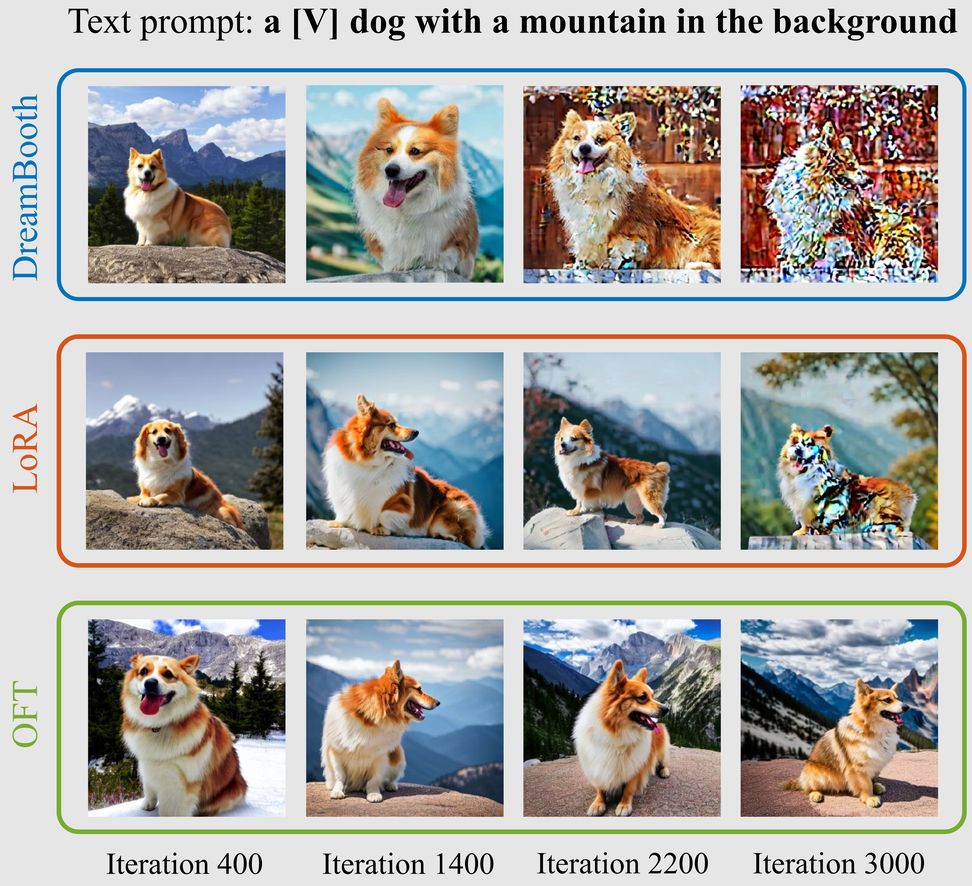
Diffusion models papers reading club starting soon! Please check:

Website LinkedIn

Kamil.deja@pw.edu.pl

Orthogonal fine-tuning

- Fine-tuning can lead to overfitting and destruction of the generative properties of the base model
- Learn layer-specific neurons orthogonal transformations
- OFT preserves hypersperical energy which characterizes pairwise relation between neurons



Qiu, Zeju, et al. "Controlling text-to-image diffusion by orthogonal finetuning." *Advances in Neural Information Processing Systems* 36 (2023): 79320-79362.

Bonus - Weigths2Weights

- Train 60 000 LoRAs of individual visual identities of celebrities
- Cast adapter weights to *w2w* space with dimensionality reduction
- Create new identities, or edit the existing ones in low-dimensional space