TIME

Text-to-Image Models for Counterfactual Explanations: a Black-Box Approach

Guillaume Jeanneret Loıc Simon Frederic Jurie

Tymoteusz Kwieciński
MI2.AI Research Seminars, summer 2024

Outline

1. Main paper contributions
2. Introduction – LDM and CEs
3. Textual inversion – how LDM can learn new concepts?
4. EDICT – perfect inversion of images
5. Description of TIME method
6. Results

Article main contributions

- A **black box** approach to generate Counterfactual Explanations
- Method uses Latent Diffusion Models
- Does **not require any optimization loop** during inference
- Very **short** and **low-resource** inference

Method	Model	Training	Specificity	Includes optimization
DiVE	VAE	Days Only	DNN	Yes
STEEX	GAN	Days Only	DNN	Yes
DIME	DDPM	Days Only	DNN	Yes
ACE	DDPM	Days Only	DNN	Yes
TIME	Text2Image	Hours	Black-Box	No

A comparison of different methods for generating Counterfactual Explanations for image classification

Latent Diffusion Models

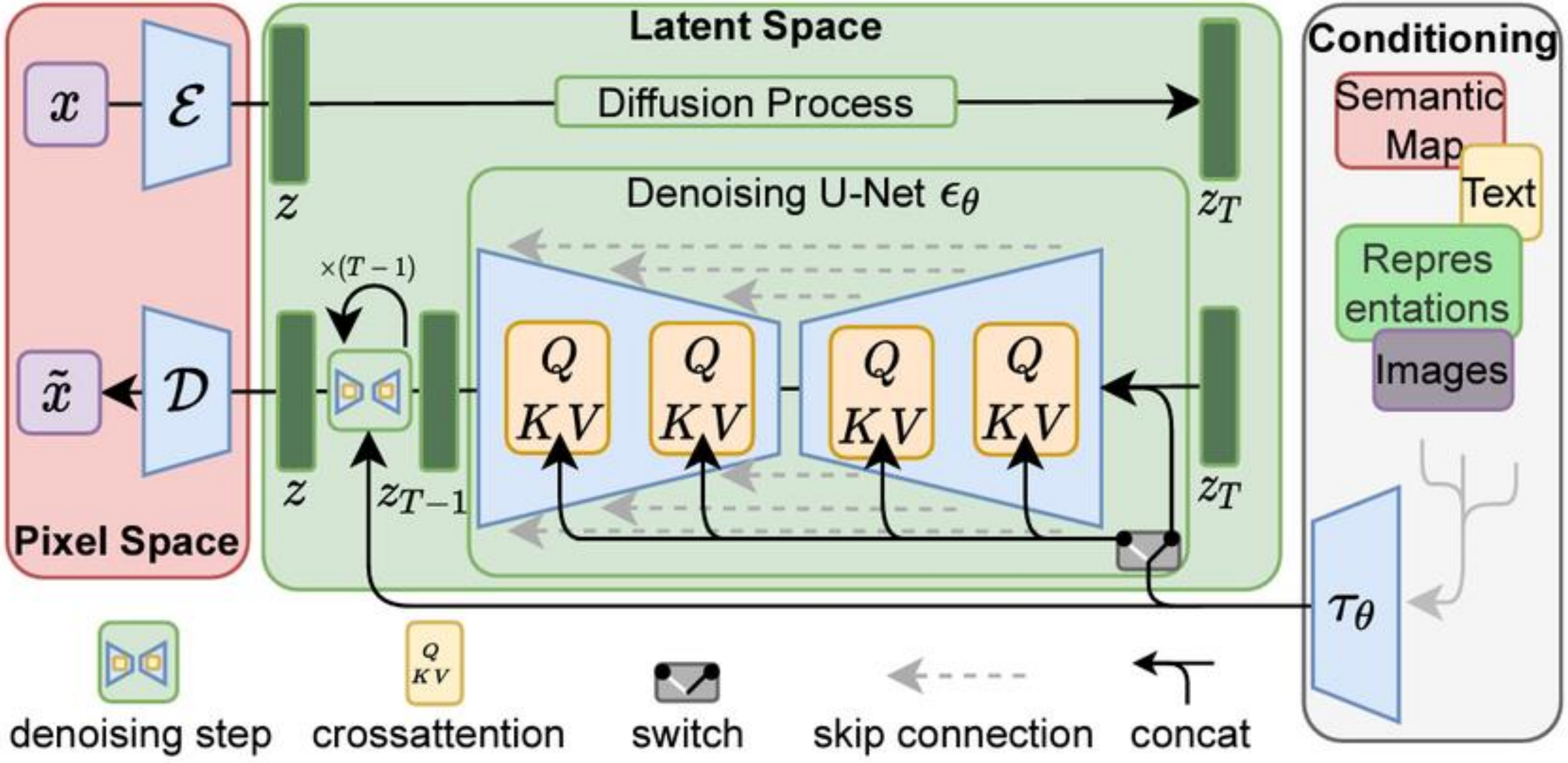
Model consists of two parts:

1. **Autoencoder**: encoder $\mathcal{E}$ trained to encode image x into latent space $\mathcal{E}\,(\mathrm{x}) = z$ and decoder D, trained to map latent back to image: $\mathrm{D}(\mathcal{E}(\mathrm{x})) \approx \mathrm{x}$
2. **Diffusion model**: model ϵ_θ iteratively denosies the noised latent, minimizing the loss in each step:

$$L_{LDM} \coloneqq E_{z\sim\mathcal{E}(x),y,\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(z_t), t, c_\theta(y)||^2]$$

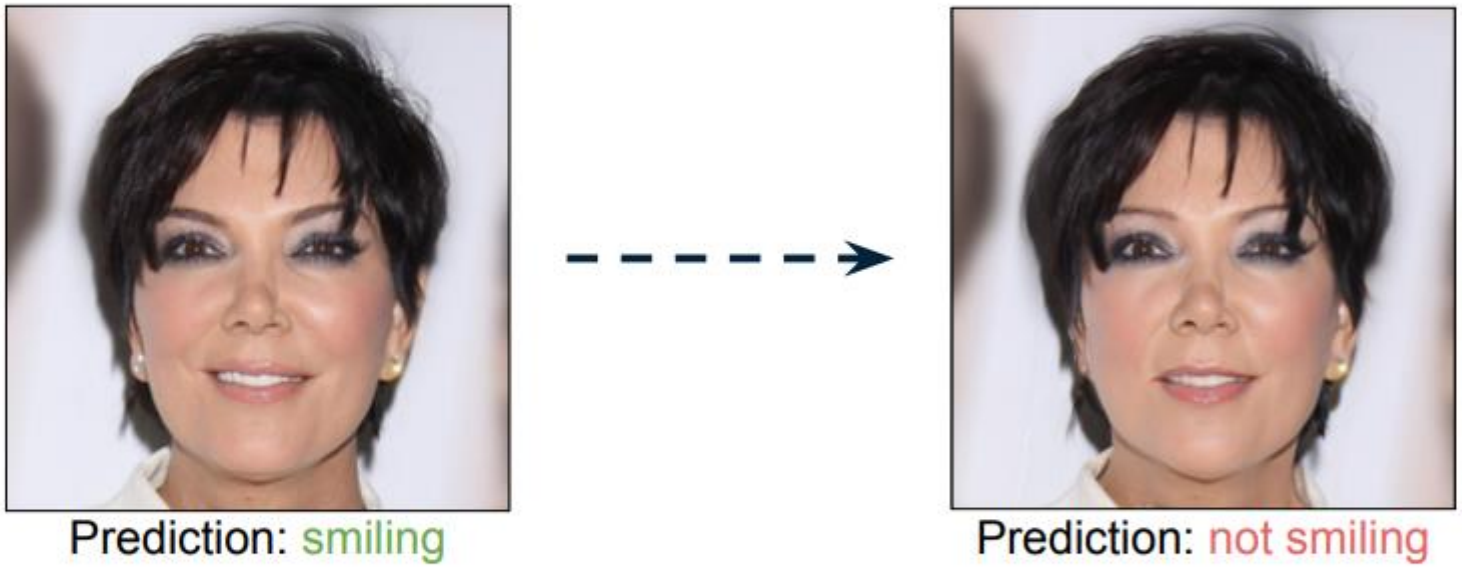
x_t is an input noisy image, t the current step and $c_\theta(y)$ textual conditioning

Latent Diffusion Model



Counterfactual Explanations for image classification

For a given **classifier** and an **image**, counterfactual explanation is an image that has a **minimal semantic change** and **flips the model's decision**



Textual inversion

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

How model can learn new concepts?

Existing method	Drawbacks
Retraining with larger dataset	Very expensive
Finetuning	Prone to forgetting the prior knowledge
Freezing the model and learning a transformation on it	Cannot utilize new concepts with the prior ones

Inversion and reconstruction of an image

- Inversion – process of finding a corresponding *latent representation* of a given image
- Reconstruction – process of finding an *image* from latent representation

Textual inversion

- Process of finding new words in the textual embedding space
- New words are denoted as S_* and correspond to learned embeddings that relates to a new concept

Method main ideas

- **Don’t change the model** that much – it may cause forgetting
- Latent representation is **expressive** enough to contain basic semantics information
- Method is based on LDM with BERT conditioning
- The only part of model that is trained are the new embeddings

Textual inversion process

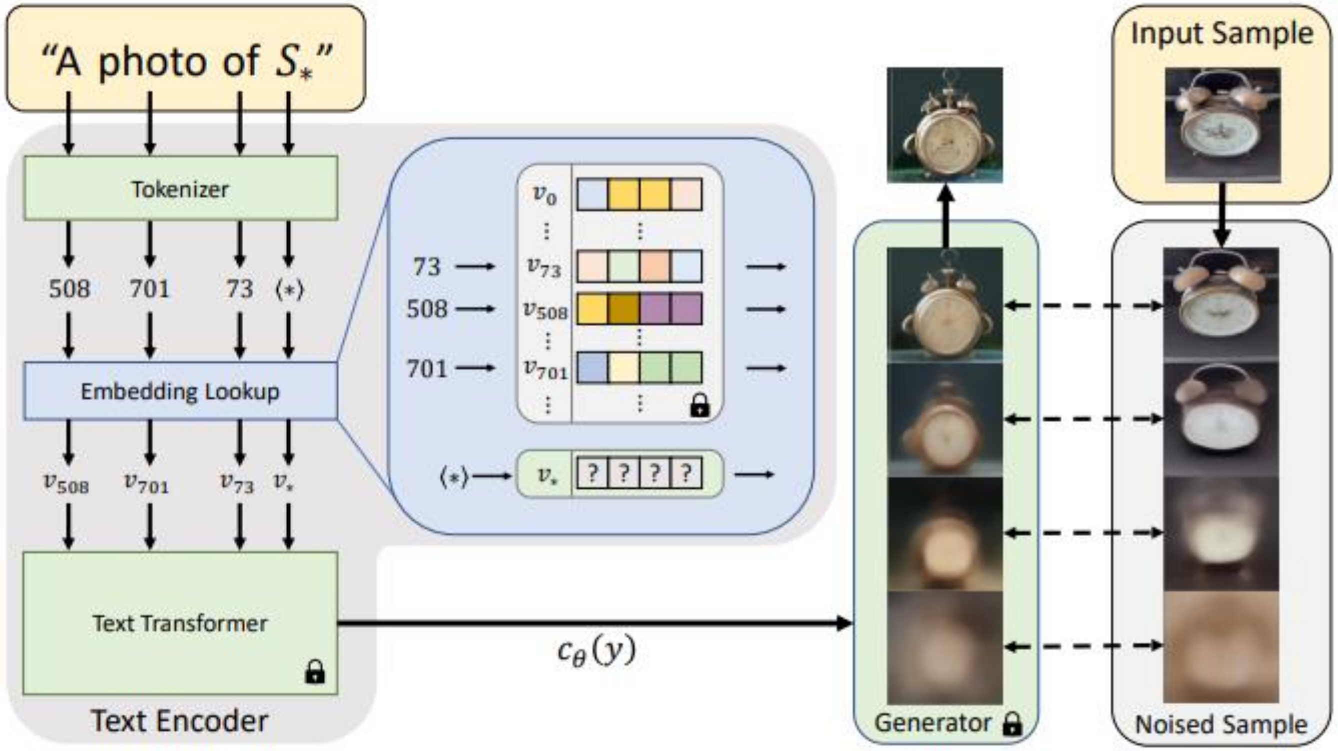
Use set **small set of images** (3-5) presenting the concept

New concept embedding v_* of pseudo-word S_* is found through **direct optimization**

$$v_* = arg\min_{v} E_{z\sim\mathcal{E}(x),y,\epsilon\sim N(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))||^2\right]$$

Reconstruction task – based on the simple prompt containing the *pseudo-word* we generate the image to be as similar to the original ones as possible

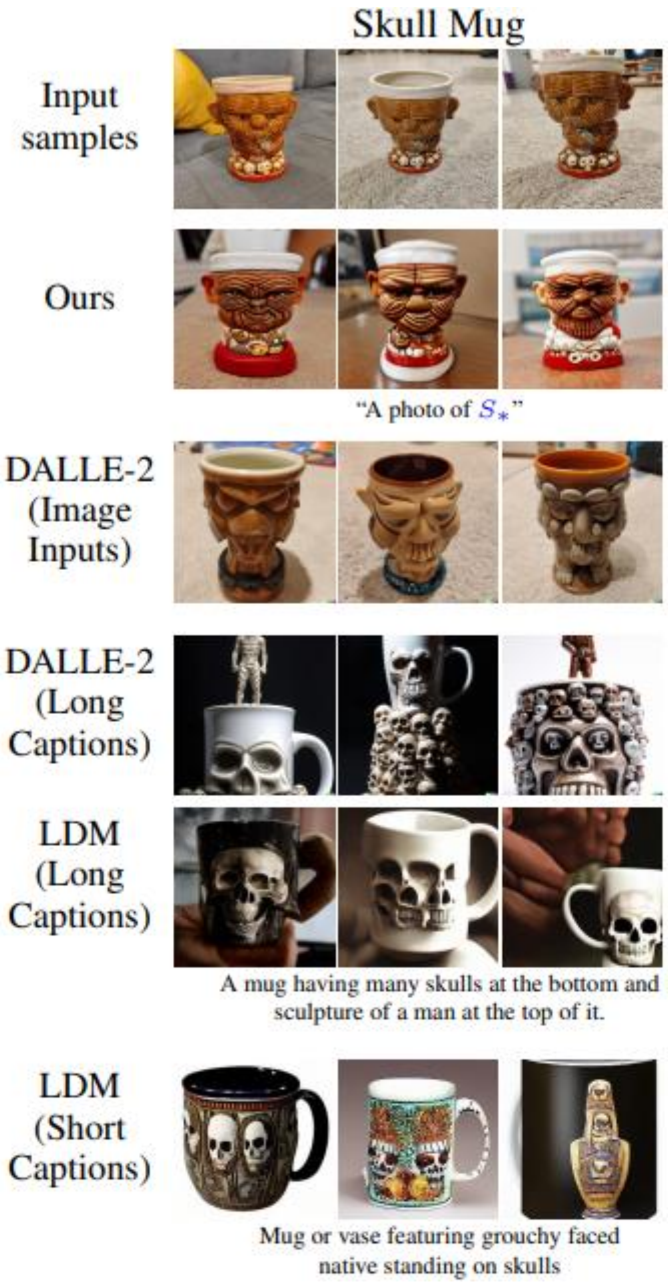
Textual inversion process



Visualization of textual inversion. The embedding v_* of token S_* is found by optimization. The generated image from the prompt should be similar to the ones from the given set

Results

- Method yields better results than textual image captioning
- Method was compared to DALLE-2 guided by image or prompt and LDM guided by longer and shorter text captions
- New embeddings are semantically meaningful

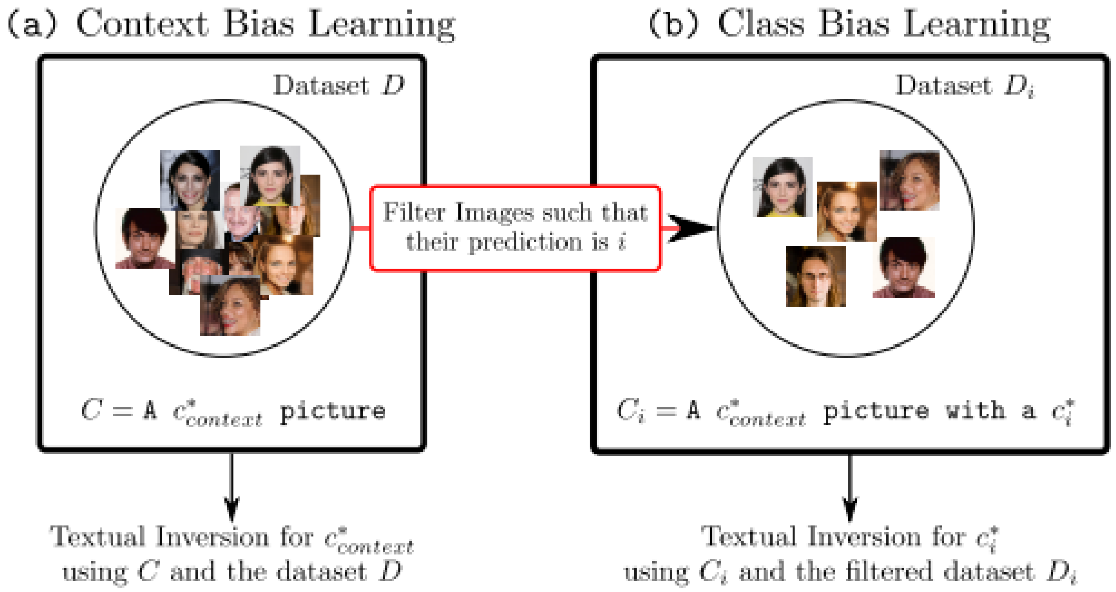


TIME

Each dataset consists of biases:

- Context bias - bias for whole dataset
- Class bias – bias of certain class

We extract the class bias for each category from dataset and create *pseudo-words* for these concepts



EDICT

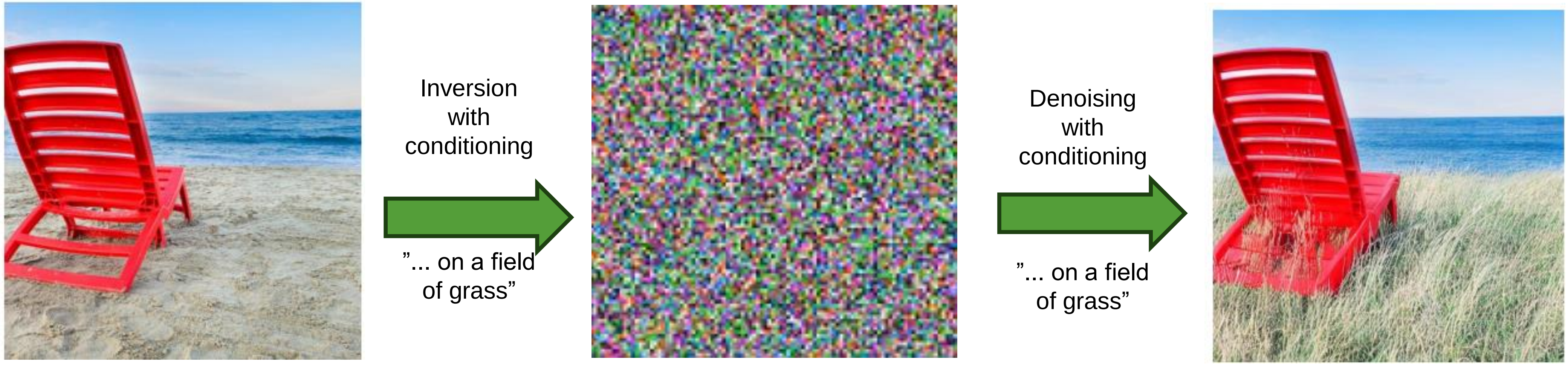
Exact Diffusion Inversion via Coupled Transformations

Inversion of images with conditioning

Having the embedded new concept, we want to **generate the concept into an image**

We **invert** the image and **denoise** the latent with conditioning

The change in the image should be **minimal**



DDIM

The denoising proces in the DDIM is **deterministic**

Reconstruction from the noised image (latent) is **exact**

The noising proces is done according to the schedule $\{\alpha_t\}_{t=0}^{T}, \alpha_T = 0, \alpha_0 = 1$

$$x_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon$$

Where x is an original image and $\epsilon \sim N(0,1)$; The **denoising** proces consists of steps:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t, t, C)}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(z_t, t, C)$$

This notation can be simplified to:

$$x_{t-1} = a_t x_t + b_t \epsilon_\theta(x_t, t, C)$$

Invertibility

Thanks to the linearity assumption we may reverse each step and obtain x_t from x_{t-1}.

$$x_t = \frac{x_{t-1} - b_t \epsilon_\theta(x_t, t, C)}{a_t} \approx \frac{x_{t-1} - b_t \epsilon_\theta(x_{t-1}, t, C)}{a_t}$$

Conditional reconstructions are **extremely disorted** and yield inconsistent results

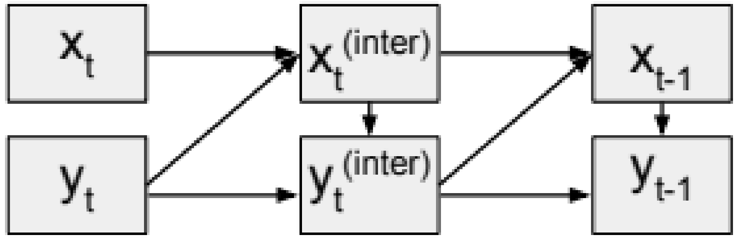
In general to properly introduce the conditioning signal, we use the Classifier-Free Guidance:

$$\epsilon_\theta(x_t, t, C) = \epsilon'_\theta(x_t, t, \emptyset) + \lambda \cdot \left(\epsilon'_\theta(x_t, t, C) - \epsilon'_\theta(x_t, t, \emptyset)\right)$$

Where ϵ'_θ is the bare network

EDICT method

- Modification of the process using with inspiration from Normalizing Flows and Ho’s method to **stabilize the inversion**
- Method reduces the impact of conditioning to preserve most of the semantic information from the original image
- It uses **two separate flows** of noising and denoising process



Information flow of EDICT

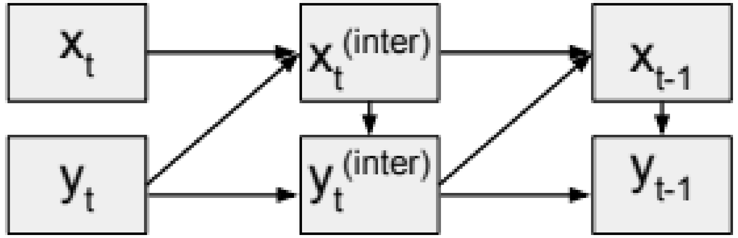
EDICT method

Denoising proces:

$$x_t^{\text{inter}} = a_t \cdot x_t + b_t \cdot \epsilon(y_t, t, C)$$
$$y_t^{\text{inter}} = a_t \cdot y_t + b_t \cdot \epsilon\left(x_t^{\text{inter}}, t, C\right)$$
$$x_{t-1} = p \cdot x_t^{\text{inter}} + (1-p) \cdot y_t^{\text{inter}}$$
$$y_{t-1} = p \cdot y_t^{\text{inter}} + (1-p) \cdot x_{t-1}$$

Deterministic inversion proces:

$$y_{t+1}^{\text{inter}} = (y_t - (1-p) \cdot x_t)/p$$
$$x_{t+1}^{\text{inter}} = \left(x_t - (1-p) \cdot y_{t+1}^{\text{inter}}\right)/p$$
$$y_{t+1} = \left(y_{t+1}^{\text{inter}} - b_{t+1} \cdot \epsilon\left(x_{t+1}^{\text{inter}}, t+1, C\right)\right)/a_{t+1}$$
$$x_{t+1} = \left(x_{t+1}^{\text{inter}} - b_{t+1} \cdot \epsilon(y_{t+1}, t+1, C)\right)/a_{t+1}$$



Information flow of EDICT

EDICT results

EDICT **doesn't require retraining** the network

Method yields much **less semantic changes** into the edited image

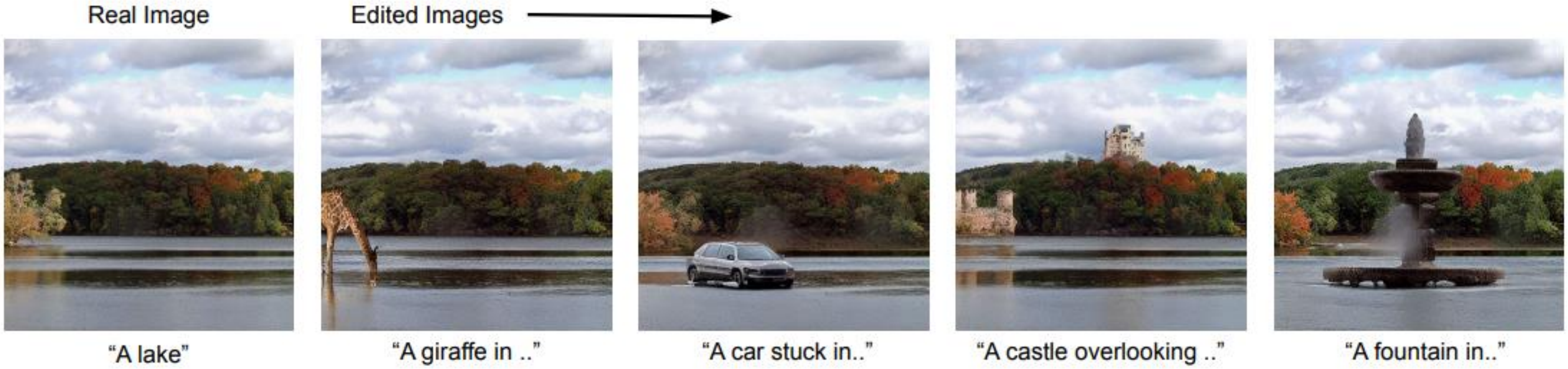
The (inter) mixing steps are required to prevent divergence of images

The double flow (x, y) **increases the inference time**,

but improves the stability of inversion

Influence of the mixing layer of steps on the inversion proces, compared to the image inverted by baseline DDIM

EDICT results



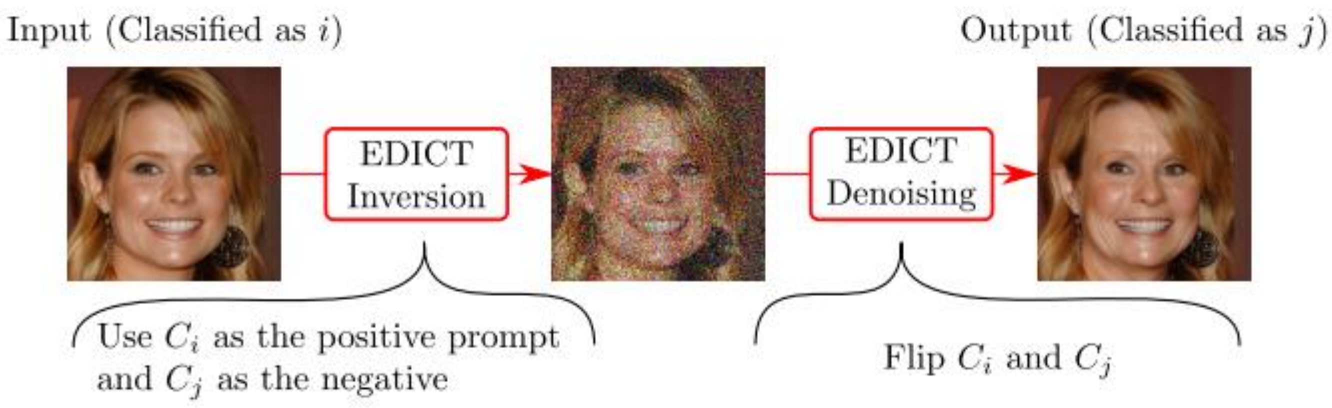
Flipping classes – counterfactual generation

- Inverting an image with a caption and denoising it with modified caption will bring semantic changes
- We use *positive* (i) and *negative* (j) *drift* terms:

$$\epsilon_\theta^c(x_t, t, C_i, C_j) = (1+\lambda)\epsilon_\theta(x_t, t, C_i) - \lambda\epsilon_\theta(x_t, t, C_j).$$

Authors also used additional hyperparameter τ – an early stop of noising process

(c) Counterfactual Generation from i to j



Quantitative assessment

Feature	Metric
Validity	Success Ratio (Flip Rate)
Sparsity and proximity (faces)	Face Similarity, MNAC
Sparsity and proximity (general purpose)	SimSiam similarity, COUT
Realism	FID, sFID
Efficiency	FLOPs

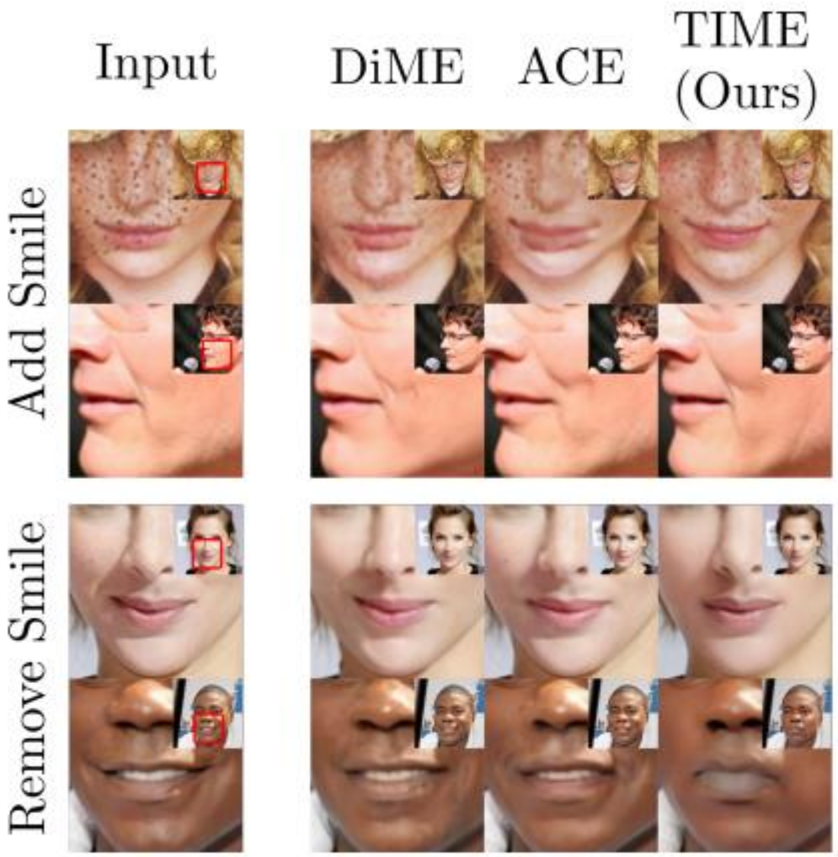
Results

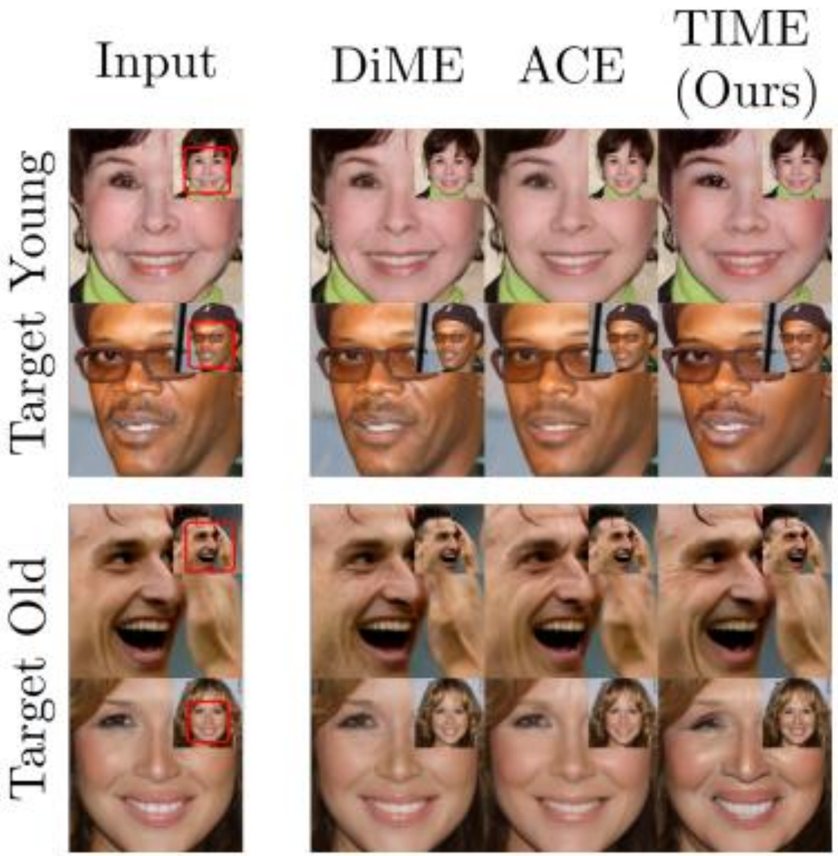
TIME yields **worse results** in every compared category

It is a **blackbox** method, with **much shorter** inference time

Method	Smile							
	FID (↓)	sFID (↓)	FVA (↑)	FS (↑)	MNAC (↓)	CD (↓)	COUT (↑)	SR (↑)
DiVE [39]	107.0	-	35.7	-	7.41	-	-	-
STEEX [23]	21.9	-	97.6	-	5.27	-	-	-
DiME [24]	18.1	27.7	96.7	0.6729	2.63	1.82	0.6495	97.0
ACE* ℓ_1 [25]	26.1	36.8	99.9	0.8020	2.33	2.49	0.4716	95.7
ACE ℓ_1 [25]	3.21	20.2	100.0	0.8941	1.56	2.61	0.5496	95.0
ACE* ℓ_2 [25]	26.0	35.2	99.9	0.8010	2.39	2.40	0.5048	97.9
ACE ℓ_2 [25]	6.93	22.0	100.0	0.8440	1.87	2.21	0.5946	95.0
TIME (Ours)	10.98	23.8	96.6	0.7896	2.97	2.32	0.6303	97.1

Visual examples





Visual examples

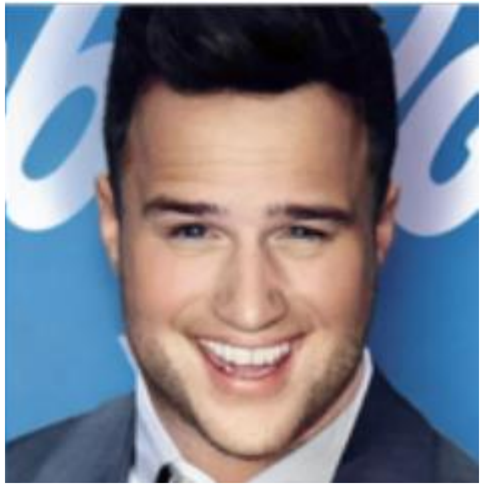
Limitations

- TIME modifications are very large in more complex cases
- On BDD100K dataset in predicting car behaviour it changes the scene a lot

Method	FID (↓)	sFID (↓)	S^3 (↑)	COUT (↑)	SR (↑)
STEEX	58.8	-	-	-	99.5
DiME	7.94	11.40	0.9463	0.2435	90.5
ACE ℓ_1	1.02	6.25	0.9970	0.7451	99.9
ACE ℓ_2	1.56	6.53	0.9946	0.7875	99.9
TIME (Ours)	51.5	76.18	0.7651	0.1490	81.8

Thank you for your attention

you at the beggining of
this presentation



you at the end of this
presentation

References

- Jeanneret et al., Text-to-Image Models for Counterfactual Explanations: a Black-Box Approach, 2024 IEEE/CVF
- Gal et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, ICLR 2023
- Wallace et al., EDICT: Exact Diffusion Inversion via Coupled Transformations, CVPR 2023
- Song et al., Denoising Dffusion Implicit Models, ICLR 2021
- Dhariwal et al. Diffusion Models Beat GANs on Image Synthesis, NeurIPS 2021
- Ulhaq et al., Efficient Diffusion Models for Vision: A Survey, IEEE TPAMI 2023