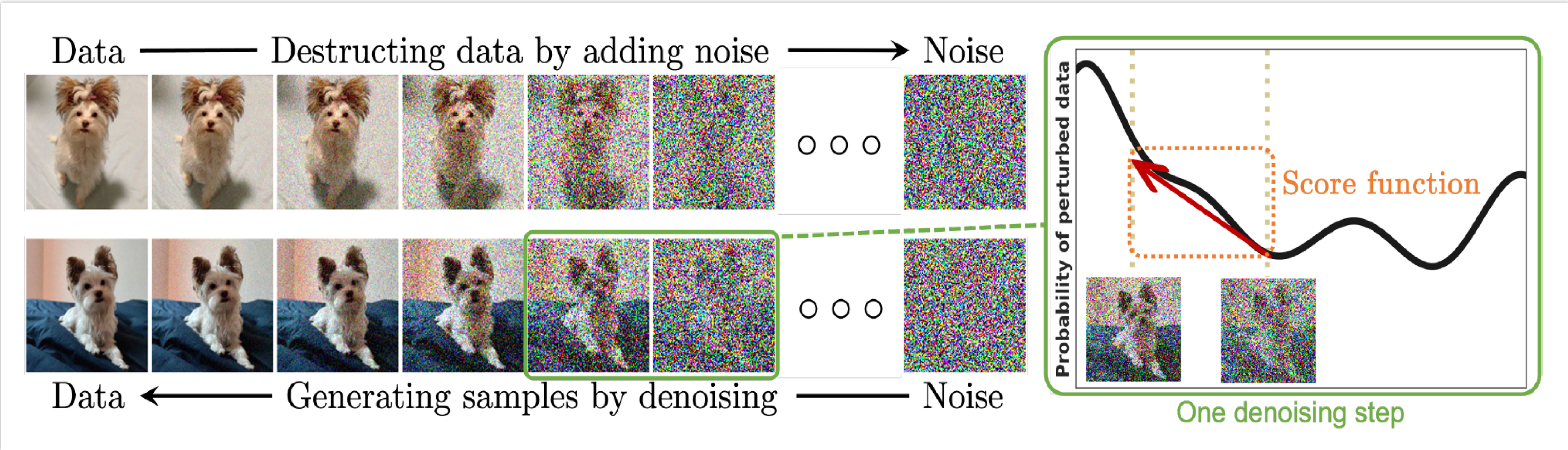
Text-guide Zero-Shot Face Multi-attribute Using Diffusion Model AutoEncoder

Dongxu Yue[1]
[1]Peking University

Motivation

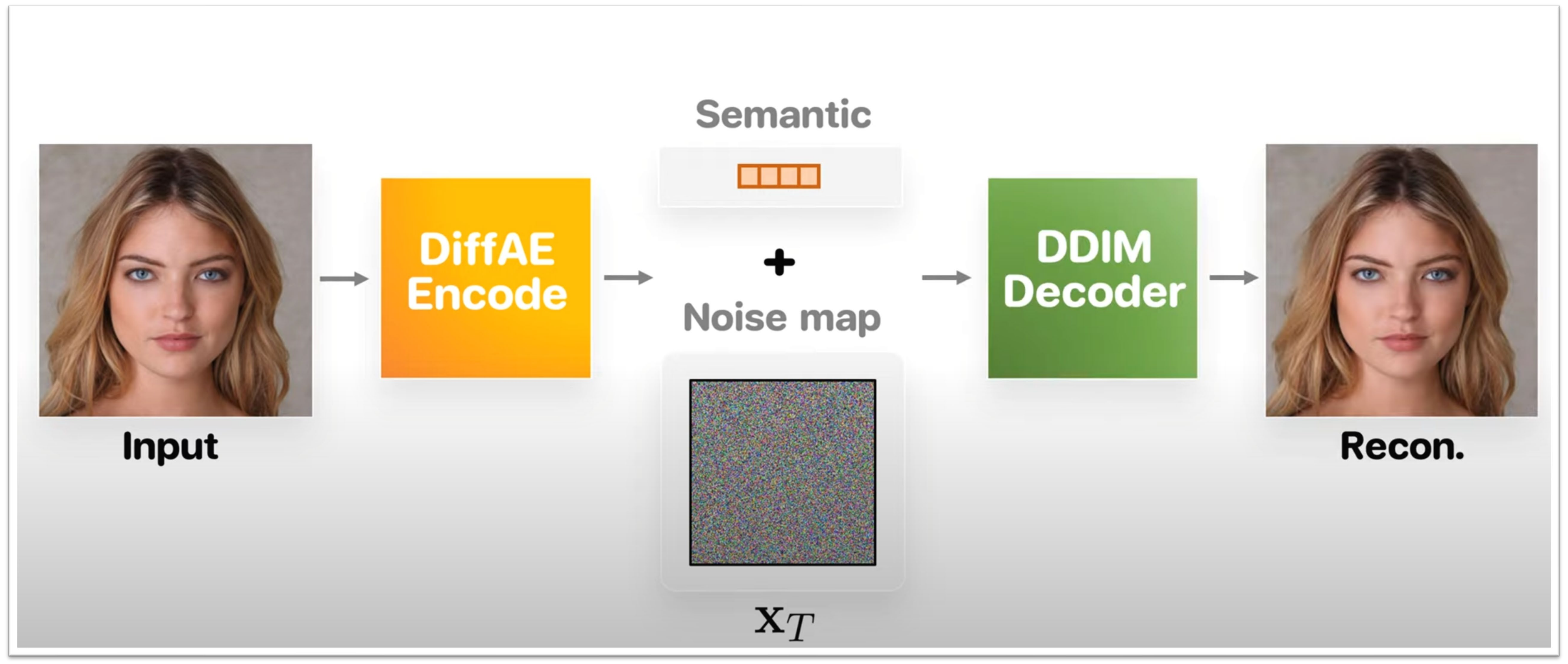


The diffusion model has powerful image generation and modification capabilities, comprehensively outperforming GAN-based approaches in image generation tasks. Recently, some methods based on GAN and CLIP have emerged for zero-shot image manipulation, but these methods have poor realizations and poor reconstructions when manipulating real images. Therefore, we propose a diffusion autoencoder-based method that enables attribute manipulation with text.

Our contributions:

- we propose a novel diffusion based manipulation method - a CLIP-guided robust zero-shot method operating in latent space.
- We have designed three different manipulation schemes with their own advantages and disadvantages.

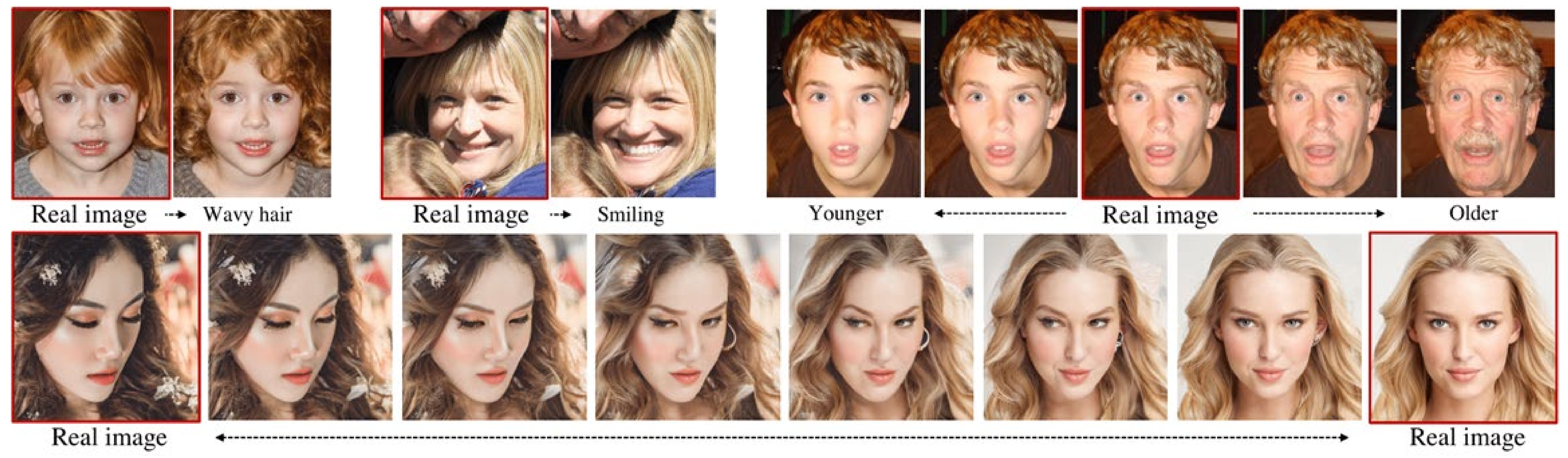
Method



Diffusion Autoencoders Architecture

Encode any image into a two-part latent code where the first part is semantically meaningful and linear, and the second part captures stochastic details, allowing near-exact reconstruction.

Attribute manipulation in hidden space



Attribute manipulation with linear classifier

By finding such a direction from the weight vector of a linear classifier trained on latent codes of negative and positive images of a target attribute consequently changes the semantic attribute in the image.

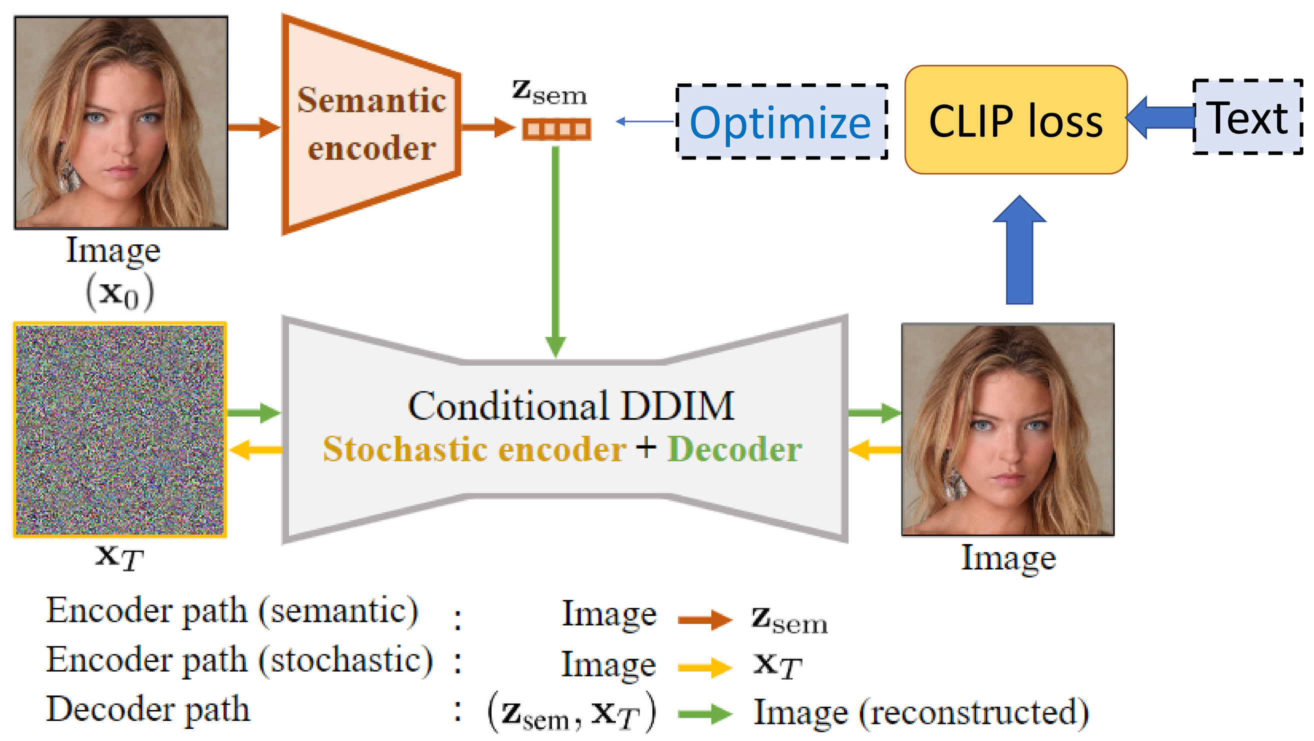
State-of-the-art performance

Image manipulation FID score

Mode	Model	Male	Smiling	Wavy Hair	Young	Blond Hair
Positive vs negative		95.82	11.15	25.04	36.75	39.65
Manipulated vs. positive	**Ours**	52.85	**9.19**	**20.80**	**20.68**	**33.51**
	StyleGAN-$\mathcal{W}$	**42.90**	18.52	27.10	31.15	33.89
Manipulated vs. negative	**Ours**	**23.15**	**7.25**	**4.89**	**11.81**	**6.79**
	StyleGAN-$\mathcal{W}$	66.92	22.15	20.70	31.15	27.54

➢ Strategy 1

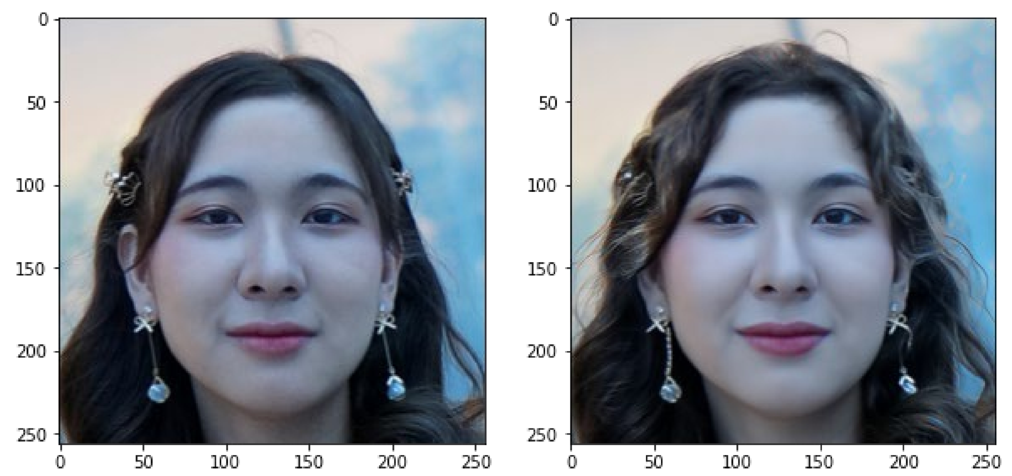
Optimize the latent space directly with CLIP-loss



➢ Strategy 2

Optimize DDIM with CLIP-loss

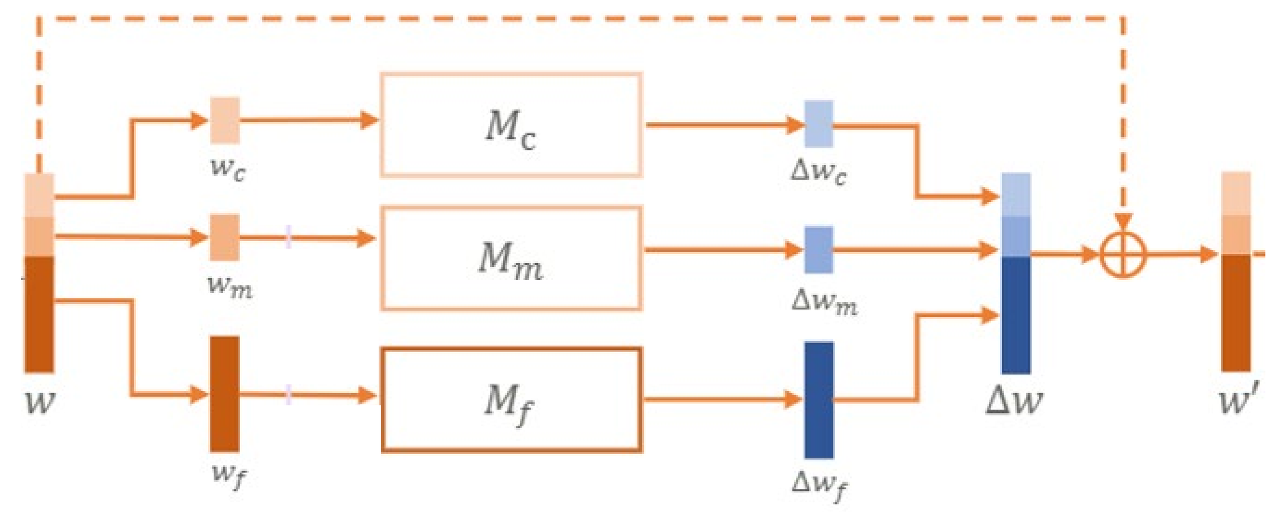
Similar to strategy 1, but instead of the hidden space, CLIP-loss optimizes the DDIM, i.e., the decoder part.



A woman with curl hair

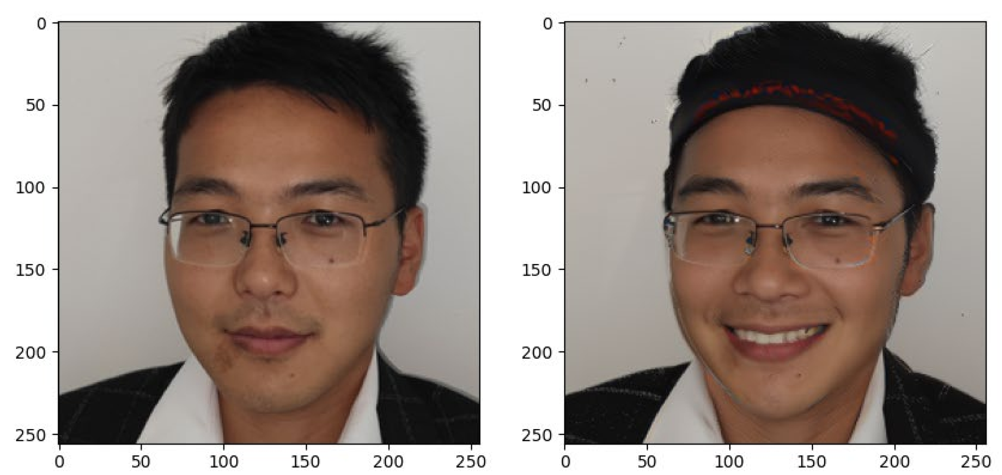
➢ Strategy 3

Design an MLP



- The inputs to MLP are latent code and Bert's encoding of Text, respectively.
- Bert was first finetuned with a single attribute and then for multi-attribute training afterwards.
- Only one training session required.

Multi-attribute manipulation



A man wearing a hat and smiling