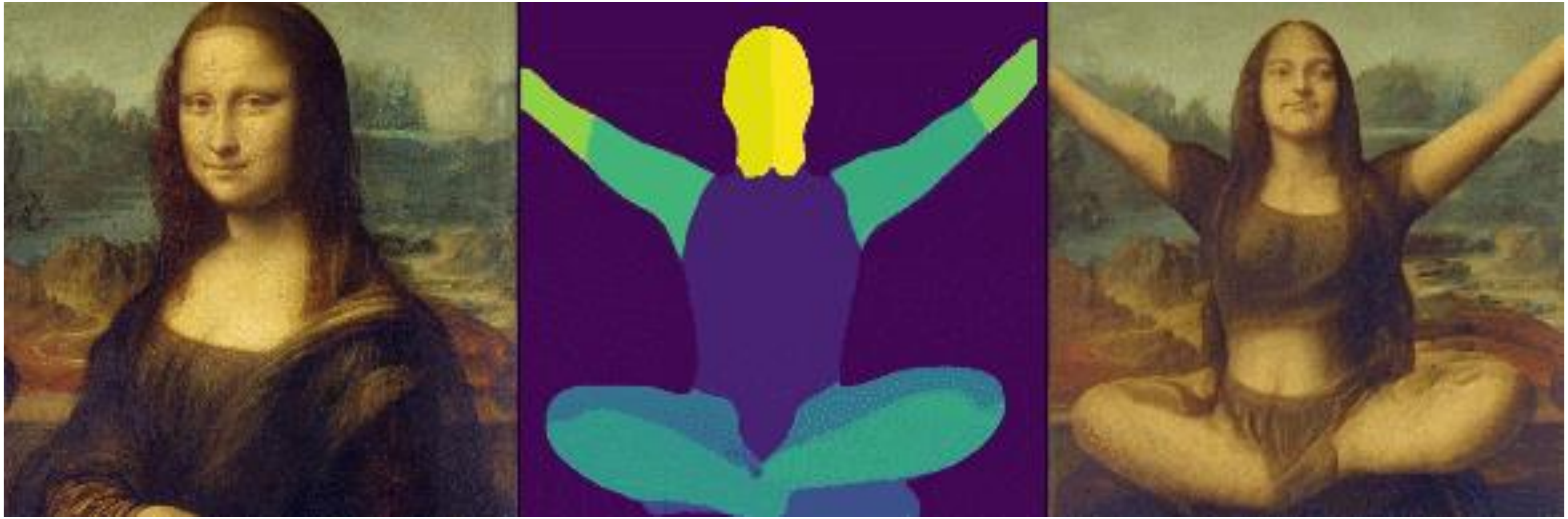
Analysis of

MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model

MagicAnimate

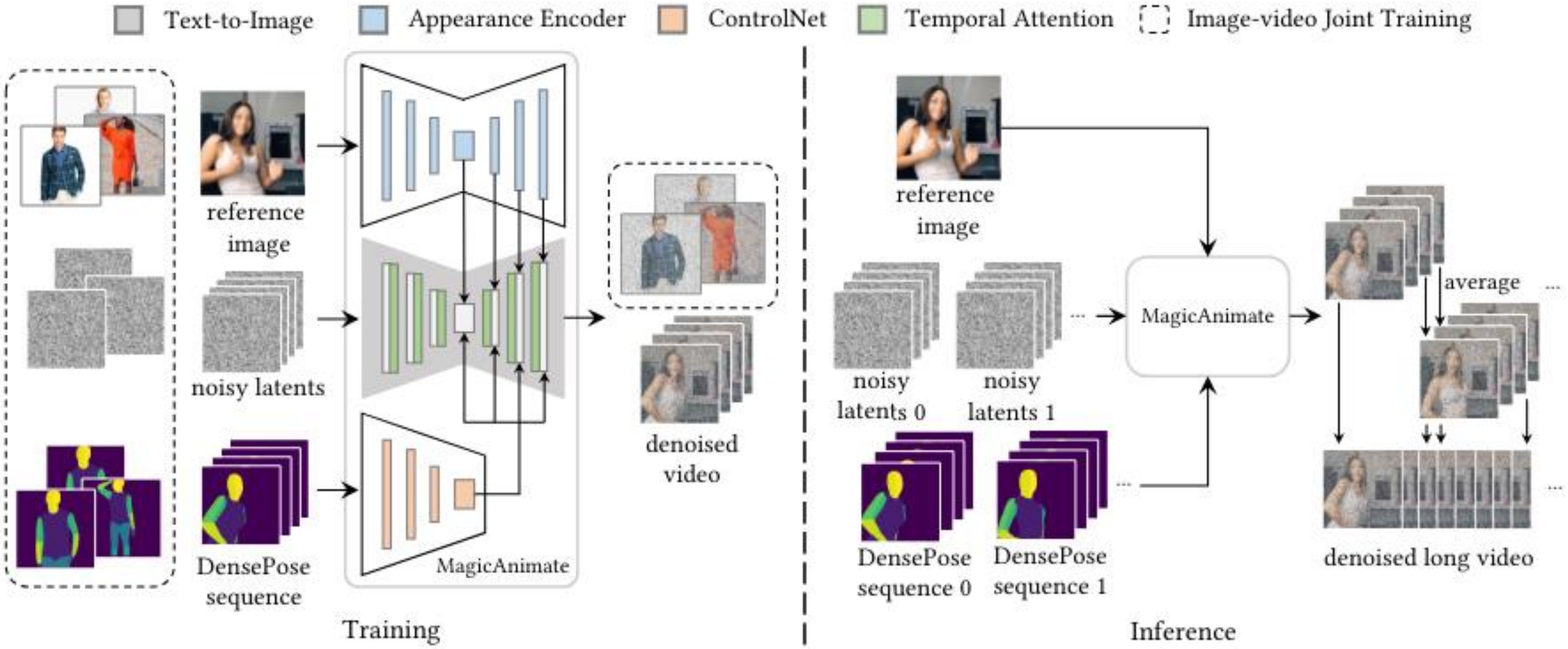
Given a sequence of motion signals such as video, depth, or pose, the image animation task aims to bring static images to life.



MagicAnimate

MagicAnimate outperforms the strongest baseline by over 38% in terms of video fidelity on the challenging TikTok dancing dataset.

MagicAnimate pipeline



Other methods’ restrictions

Other methods’ restrictions

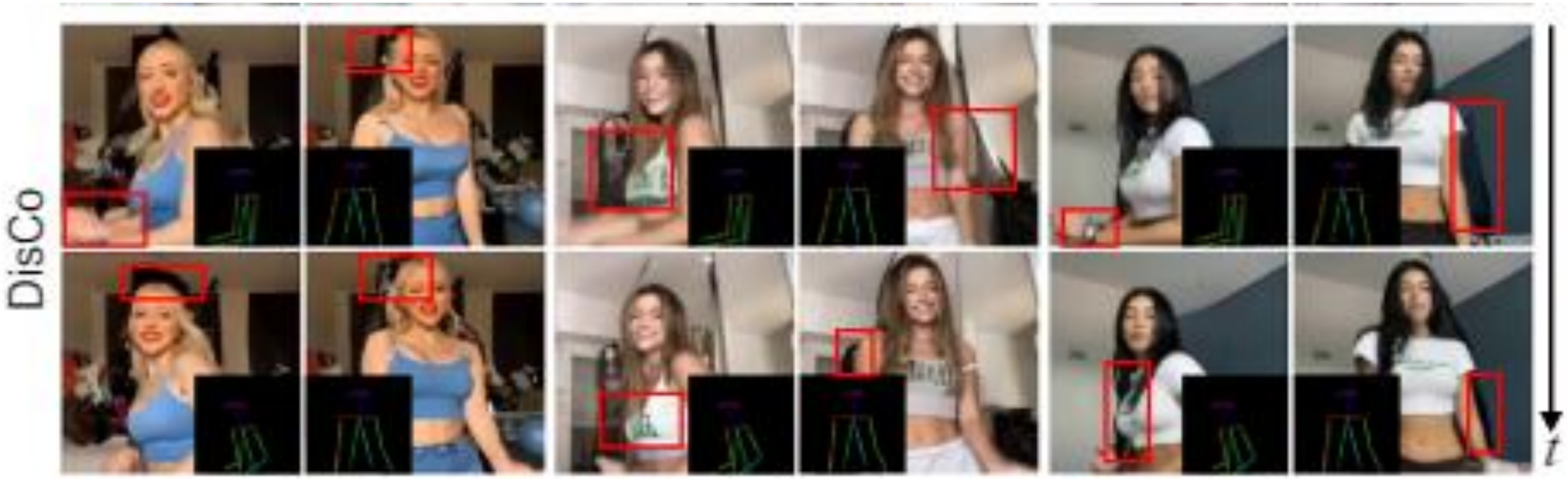
- GAN(Generative Adversarial Networks)

- Integrates the generative model and the discriminative model .The Generator creates data samples from random noise, and the Discriminator evaluates whether the samples are real or fake, using adversarial training to improve both.

- Limitations: Result in unrealistic details in occluded areas and have limited generalization across different identities.

Other methods' restrictions

- Diffusion-Based Methods: Process video frames individually and use models like CLIP, leading to flickering and less effective detail preservation.



What did MagicAnimate done?

TWO MAIN INNOVATIONS:

1. Using diffusion model
2. Novel appearance encoder

What is Diffusion Model ?

Diffusion Model

“In machine learning, diffusion models, also known as diffusion probabilistic models or score-based generative models, are a class of **latent variable generative models**. (Chang et al.)

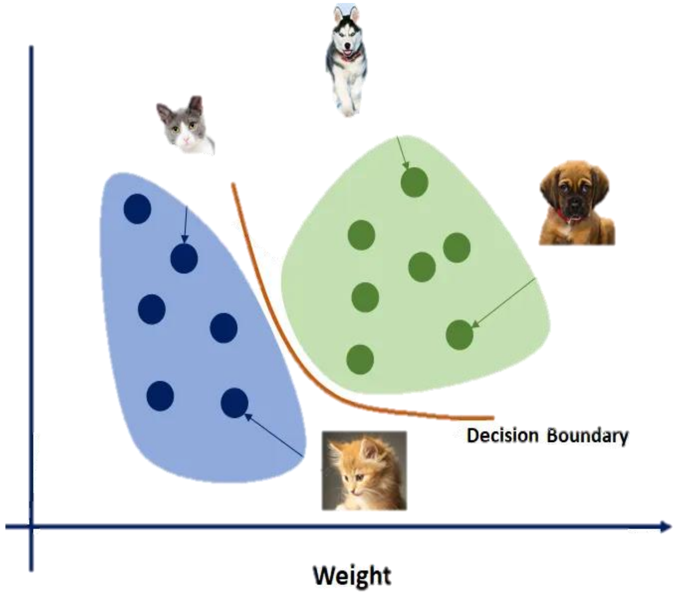
Statistical models

1.Generative model

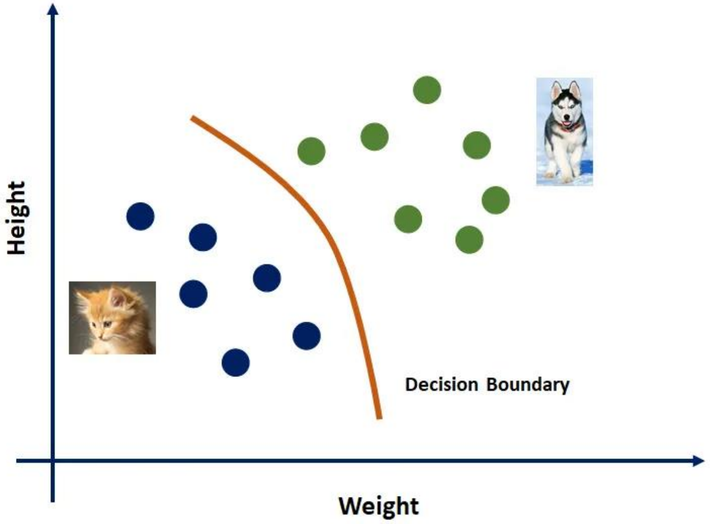
- Joint probability distribution P(X,Y)
- Observable variable(input) – X   Target variable(output) – Y
- It will observe the probability that X AND Y both exist, as it understands the relationship between X AND Y, it can generate new instances of the data

2.Discriminative model

- Conditional probability P(Y|X = x)
- It learns the probability of the target (Y) given the input (X)
- Differenciate X to different classes
- Learns the boundary between outputs, focuses on the relationship from input to output.



Generative Model



Discriminative Model

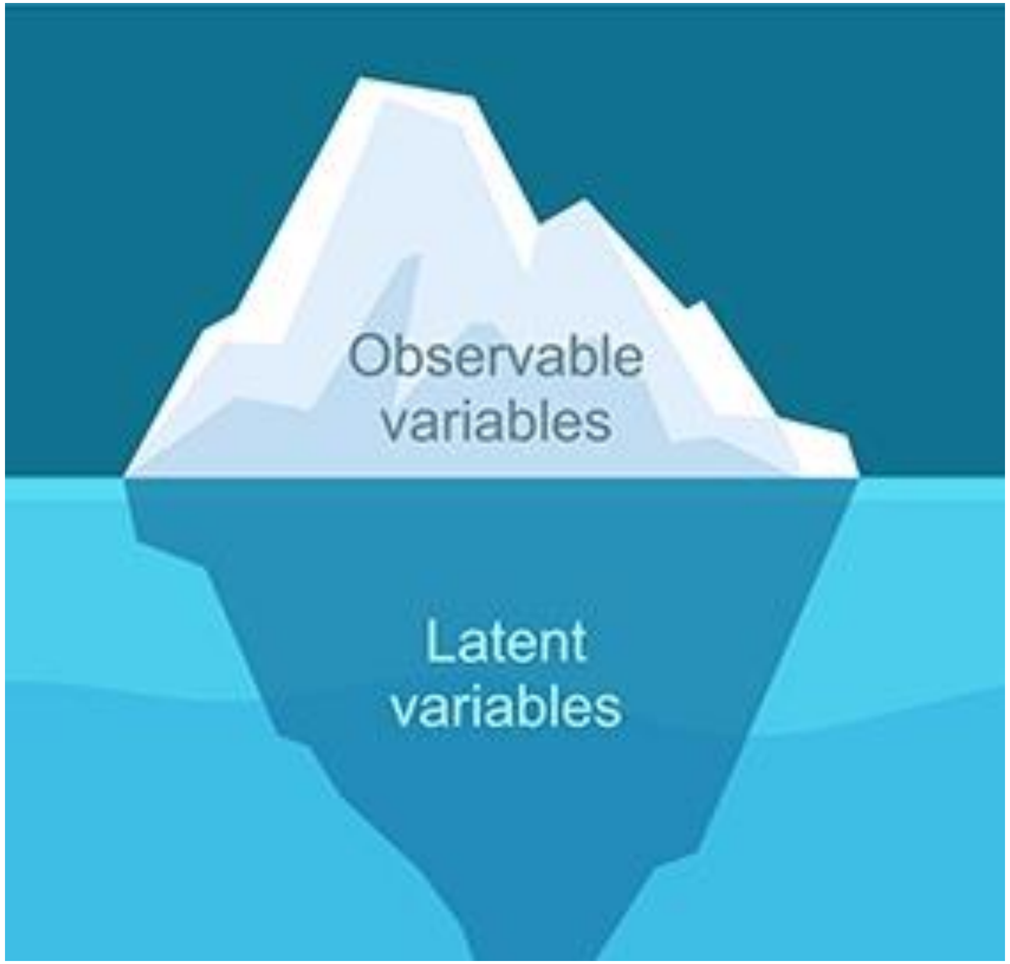
Latent and observable variable

Latent Variable:

- Hidden variables which cant be directly observed
- Represent hidden or underlying factors that influence the observable data.

Observable Variable:

- variables that can be directly measured or observed.
- Input for various statistical models



Diffusion Model

Diffusion model is a latent variable generative model. (Chang et al.)

What is a latent variable generative model?

Latent Variable Generative Model

- Generative models
- Utilize latent variables to represent the underlying structure of the observed data.
- The observed data x is generated from some latent variables z through a generative process $p(x|z)$.

Examples of Latent Variable Generative Model

- Variation autoencoder(VAE)
- Diffusion model

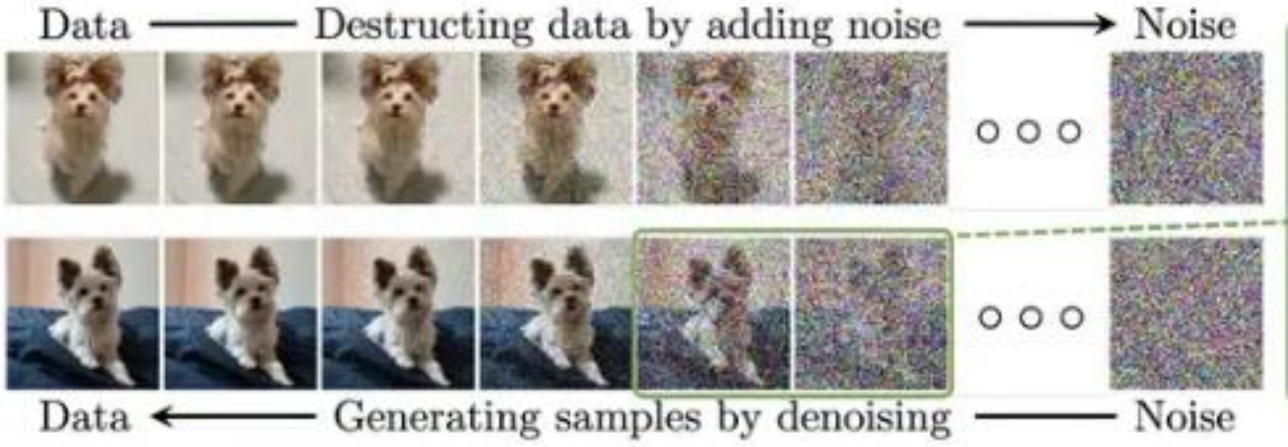
Typical diffusion model

Forward Process

1.Start with the input image

1.Add Gaussian noises in the pictures

2.The picture will turn into almost pure no


Data
Destructing data by adding noise
Noise
Data
Generating samples by denoising
Noise

Reverse Process

1.Start with the noisy images

2.Denoise the images guiding by the input motion sequence

3. The appearance encoder will retain and enhance the details of the reference identity throughout the denoising process.

4. Output

Diffusion model in MagicAnimate

Temporal Consistency Modelling

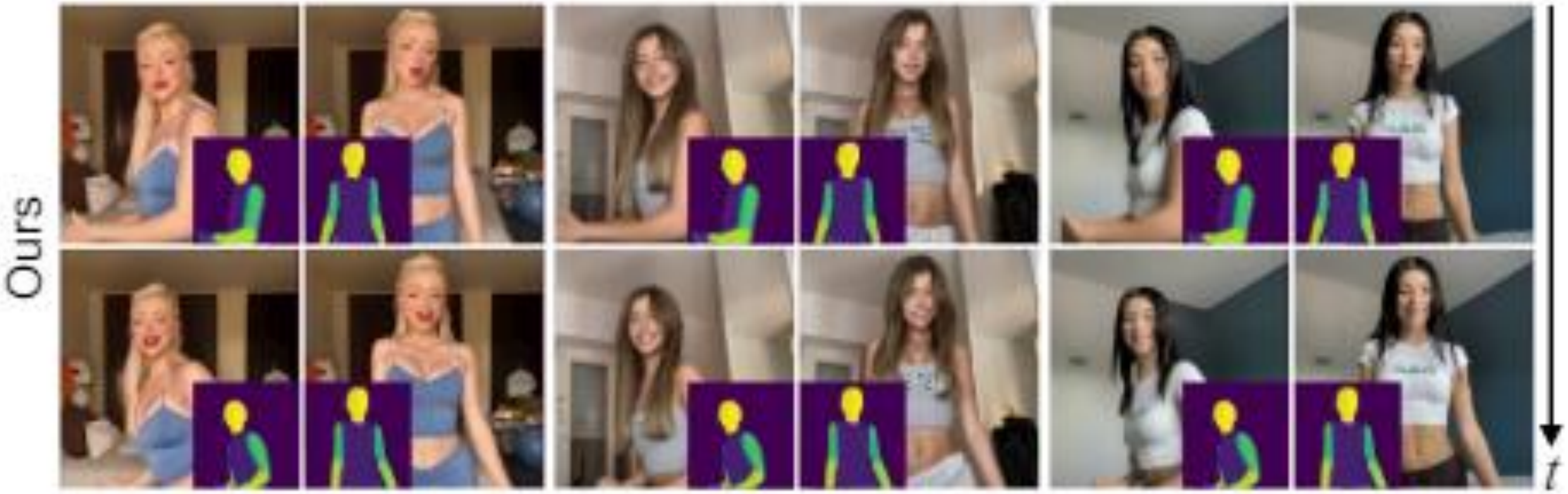
Extend the image diffusion model to the video domain

Through this, MagicAnimate aggregates temporal information from neighbouring frames and synthesizes K* frames with improved temporal consistency.

*where K is the length of the video frames

By using diffusion model

- Enhanced temporal consistency
- Preserved reference image faithfully
- Improved animation fidelity

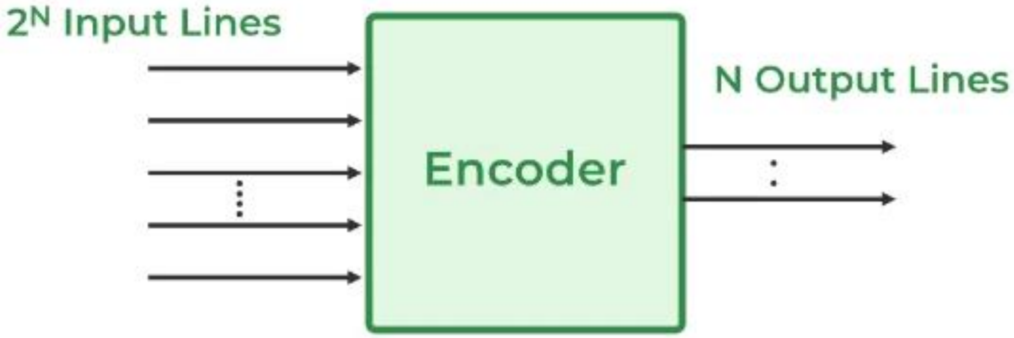


Another Main Innovation

Novel appearance encoder

Novel appearance encoder

- maintain the appearance coherence across frames
- retain the complex details of the reference image
- Main goal: Encode Iref with detailed features related to identity and background.

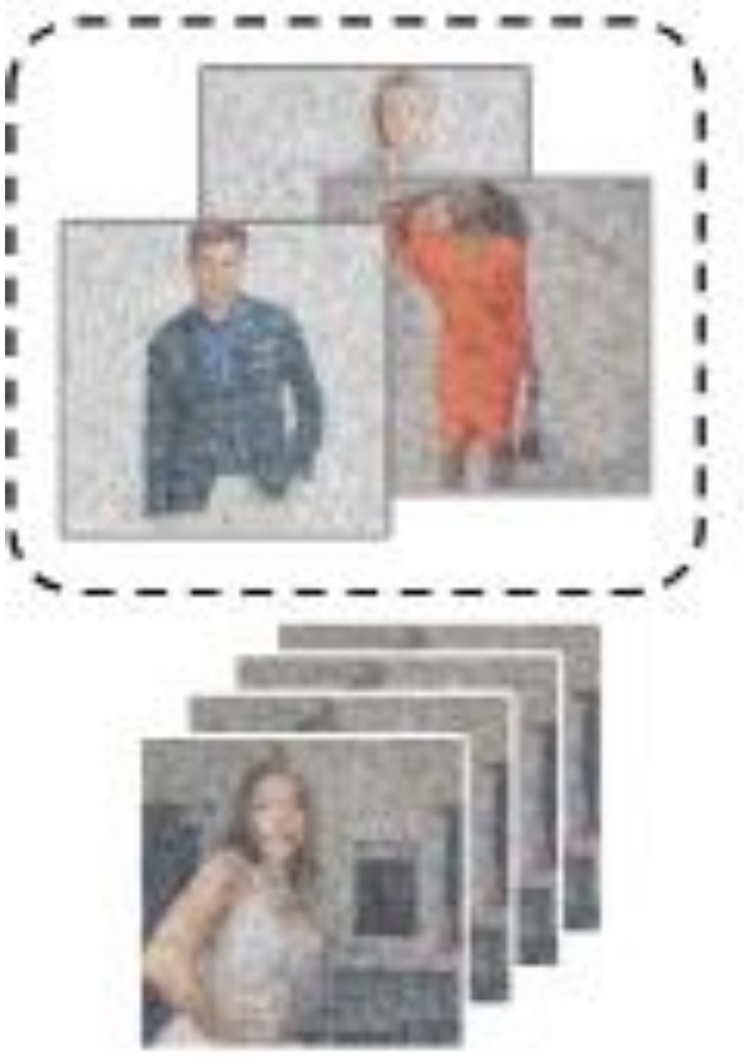


Novel appearance encoder

- 1. Extract key details such as facial features, clothing, and background.

- 2. These encoded features are concatenated with the latent noise features from the diffusion model to guide the animation process.

- 3. The integrated features are used by the diffusion model to ensure that the generated video frames maintain the detailed appearance of the reference image.



Incorporation of Temporal Consistency Modelling and Appearance Encoder

1. Enhances the preservation of identity and background details from the reference image.

2.Ensures that the generated video maintains coherence across frames.

Other two innovations

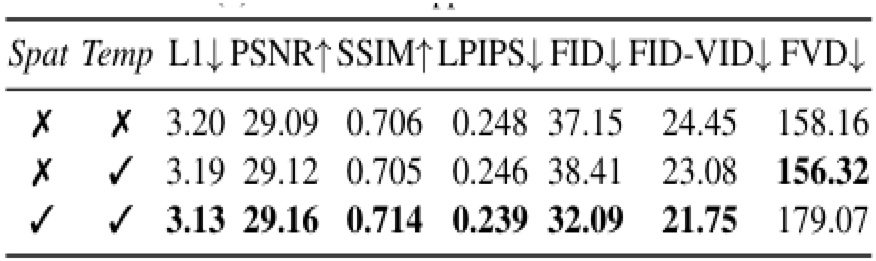
1. Image-Video Joint Training

2. Simple Video Fusion Technique

Image-Video Joint Training

By training a model using two different types of data: images and videos , the output will be handling both static images and dynamic sequences.

This process helps the model to generate high-quality, temporally consistent animations by learning from both static images and dynamic videos.

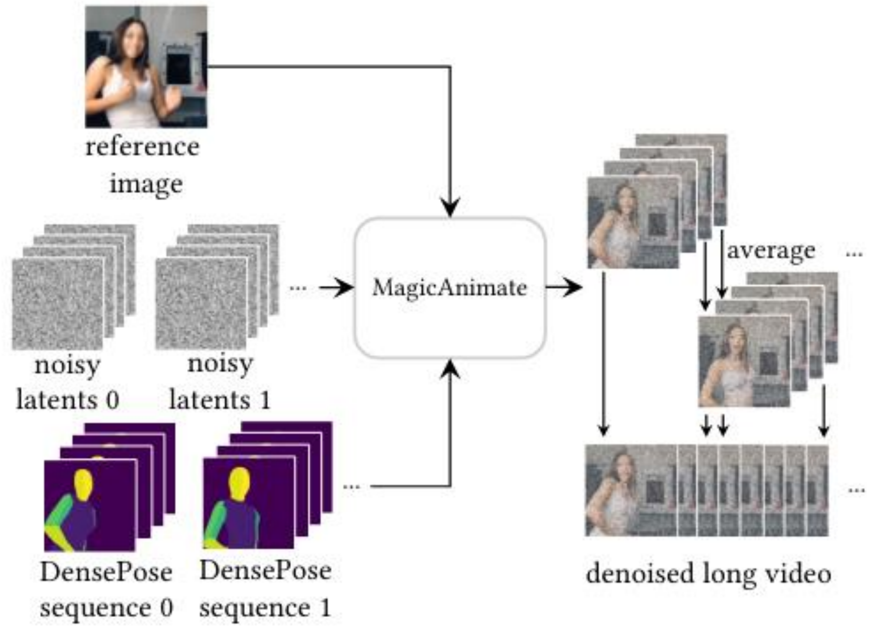


Spat	*Temp*	L1↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
✗	✗	3.20	29.09	0.706	0.248	37.15	24.45	158.16
✗	✓	3.19	29.12	0.705	0.246	38.41	23.08	**156.32**
✓	✓	**3.13**	**29.16**	**0.714**	**0.239**	**32.09**	**21.75**	179.07

(c) The effect of image-video joint training.

Simple Video Fusion Technique

By dividing the long motion sequence into multiple segments with temporal overlap, the transition smoothness will be improved



Avg	L1↓	FID↓	FID-FVD↓
w/o	3.21	32.99	22.50
w/	**3.13**	**32.08**	**21.75**

(d) The effect of the inference-stage temporal video fusion.

Noise	L1↓	FID↓	FID-FVD↓
diff	**3.03**	32.74	22.50
same	3.13	**32.08**	**21.75**

(e) The effect of sharing the same initial noises for all the video segments.

Table 2. Ablations of MagicAnimate on TikTok dataset, with best results in **bold**. We vary our architectural designs and training strategies to investigate their effectiveness. We report L1×10^{-4} for numerical simplicity.

In a nutshell MagicAnimate

- Video Diffusion Model: encodes temporal information by incorporating temporal attention blocks into the diffusion network
- Innovative Appearance Encoder: Preserves human identity, background, and details more effectively compared to other methods.
- Image-Video Joint Training: Leverages diverse single-frame image data to improve detail fidelity and modelling capability.
- Simple Video Fusion Technique: Ensures smooth transitions for long video animations.

Limitations of MagicAnimate

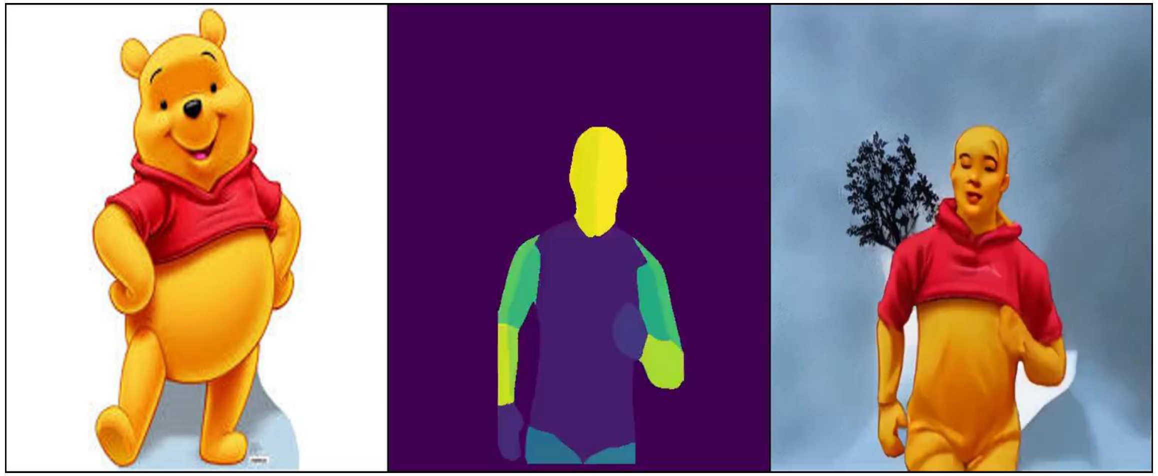
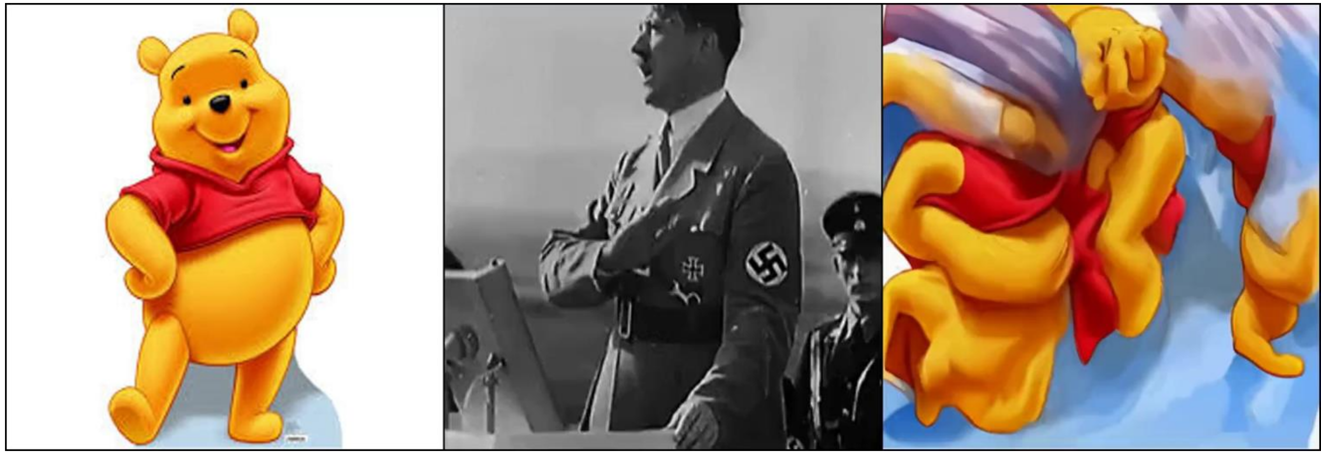
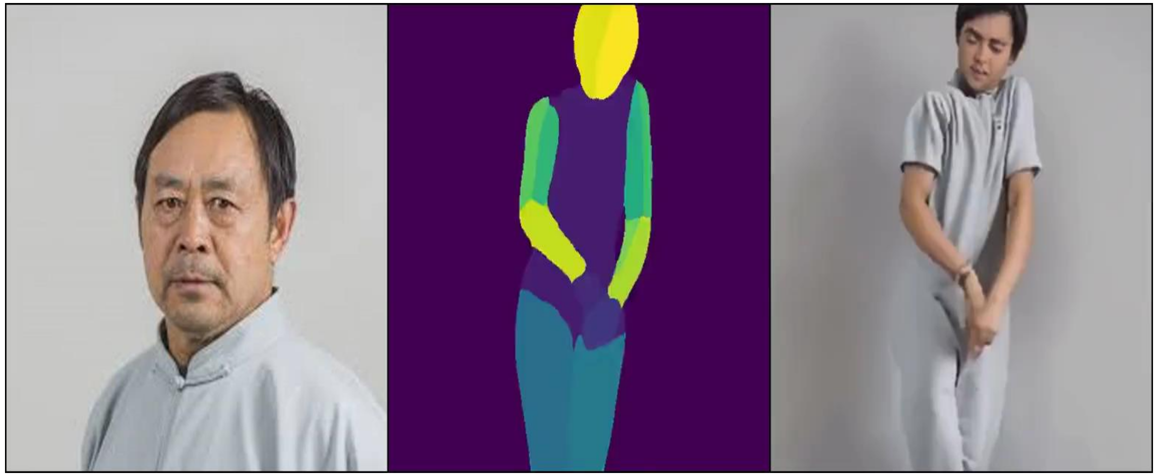
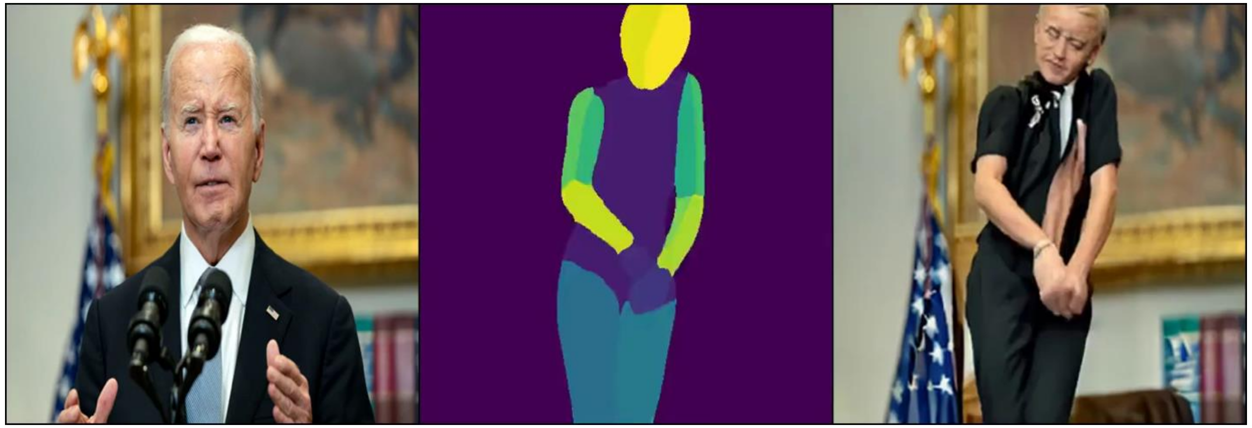
Limitations of MagicAnimate

- Long duration of processing

It needs a lot of time to process a video with a few seconds

Limitations of MagicAnimate

- Sometimes with unprecise output

Works Cited

Chang, Ziyi, et al. *On the Design Fundamentals of Diffusion Models: A Survey*. 19 Oct. 2023, arxiv.org/pdf/2306.04542. Accessed 25 June 2024

Mitchell, Tom M. *Learning Classifiers Based on Bayes Rule*. 1 Oct. 2020, www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf. Accessed 25 June 2024.

Xu, Zhongcong, et al. *MagicAnimate: Temporally Consistent Human Image Animation Using Diffusion Model*. 27 Nov. 2023, arxiv.org/pdf/2311.16498. Accessed 23 June 2024.