JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Diffusion Video Autoencoders:

Toward Temporally Consistent Face Video Editing via Disentangled Video Encoding

CVPR 2023 (TUE-PM-188)

Gyeongman Kim[1] Hajin Shim[1] Hyunsu Kim[2] Yunjey Choi[2] Junho Kim[2] Eunho Yang[1,3]

[1]KAIST [2]NAVER AI Lab [3]AITRICS

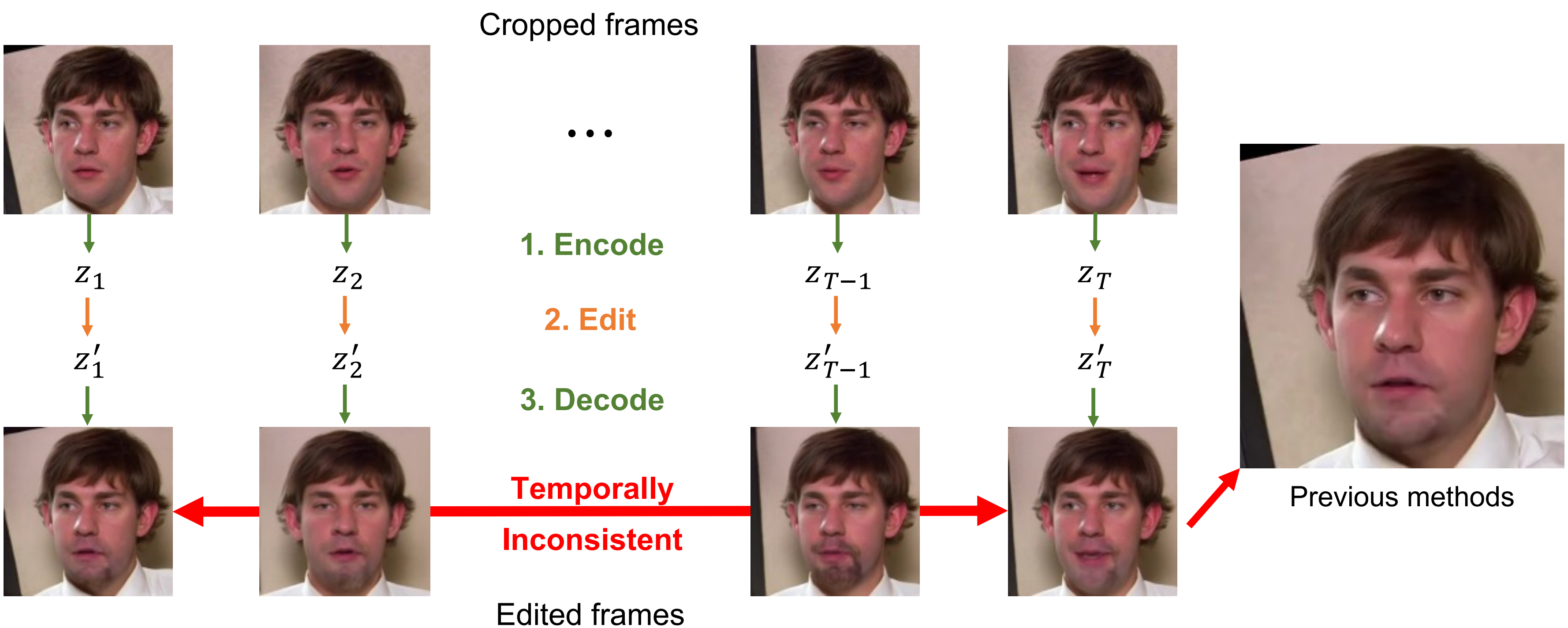
Machine Learning & Intelligence Laboratory

1 min Summary

Problem: Temporal consistency

- **Face video editing**: The task of modifying certain attributes of a face in a video
- All previous methods use GAN to edit faces for each frame **independently**

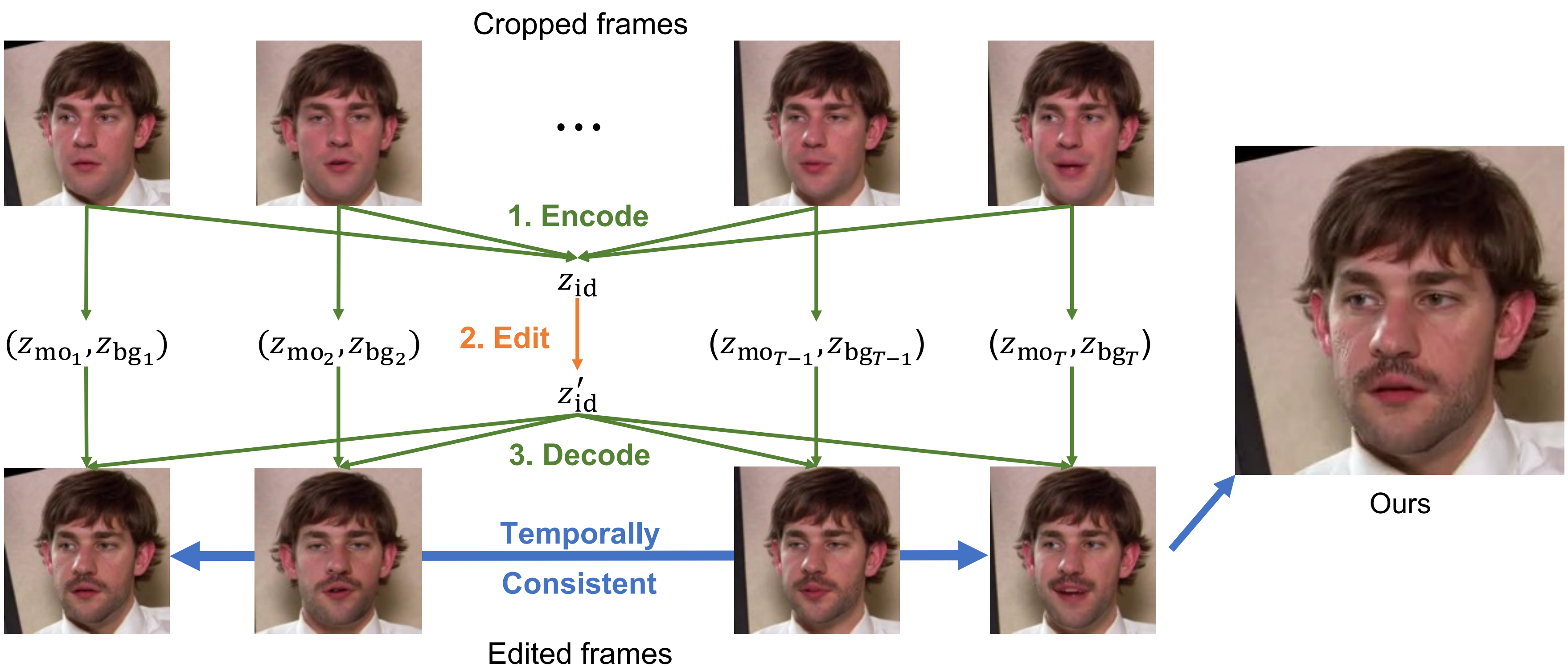
→ Modifying attributes, such as beards, causes **temporal inconsistency** problem



Solution: Decompose a video into a single identity, etc.

- **Diffusion Video Autoencoders**
  - Decompose a video into {**single identity** z_{id}, **each frame** (**motion** z_{mo_t}, **background** z_{bg_t})}
  - video → decomposed features $\left(z_{\text{id}}, \{z_{\text{mo}_t}\}_{t=1}^{T}, \{z_{\text{bg}_t}\}_{t=1}^{T}\right)$ → video

→ Entire frame can be edited **consistently** with **single modification** of the identity feature



Solution: Decompose a video into a single identity, etc.

- **Diffusion Video Autoencoders**
  - Decompose a video into {**single identity** z_{id}, **each frame** (**motion** z_{mo_t}, **background** z_{bg_t})}
  - video → decomposed features $\left(z_{\text{id}}, \{z_{\text{mo}_t}\}_{t=1}^{T}, \{z_{\text{bg}_t}\}_{t=1}^{T}\right)$ → video

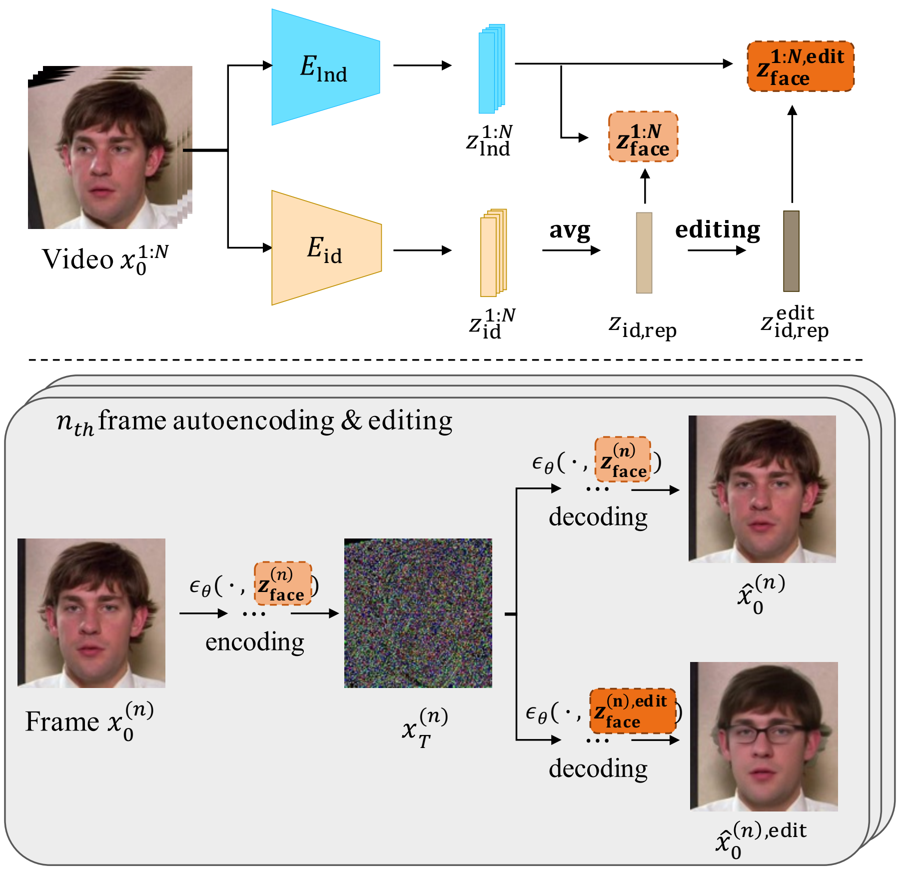
→ Entire frame can be edited **consistently** with **single modification** of the identity feature

Original | Latent Transformer (ICCV 2021) | STIT (arXiv 2022) | VideoEditGAN (ECCV 2022) | Ours

Only ours successfully produces the **temporally consistent** result!

Paper Details

Method Overview: video autoencoding & editing pipeline



- Design a diffusion video autoencoder:

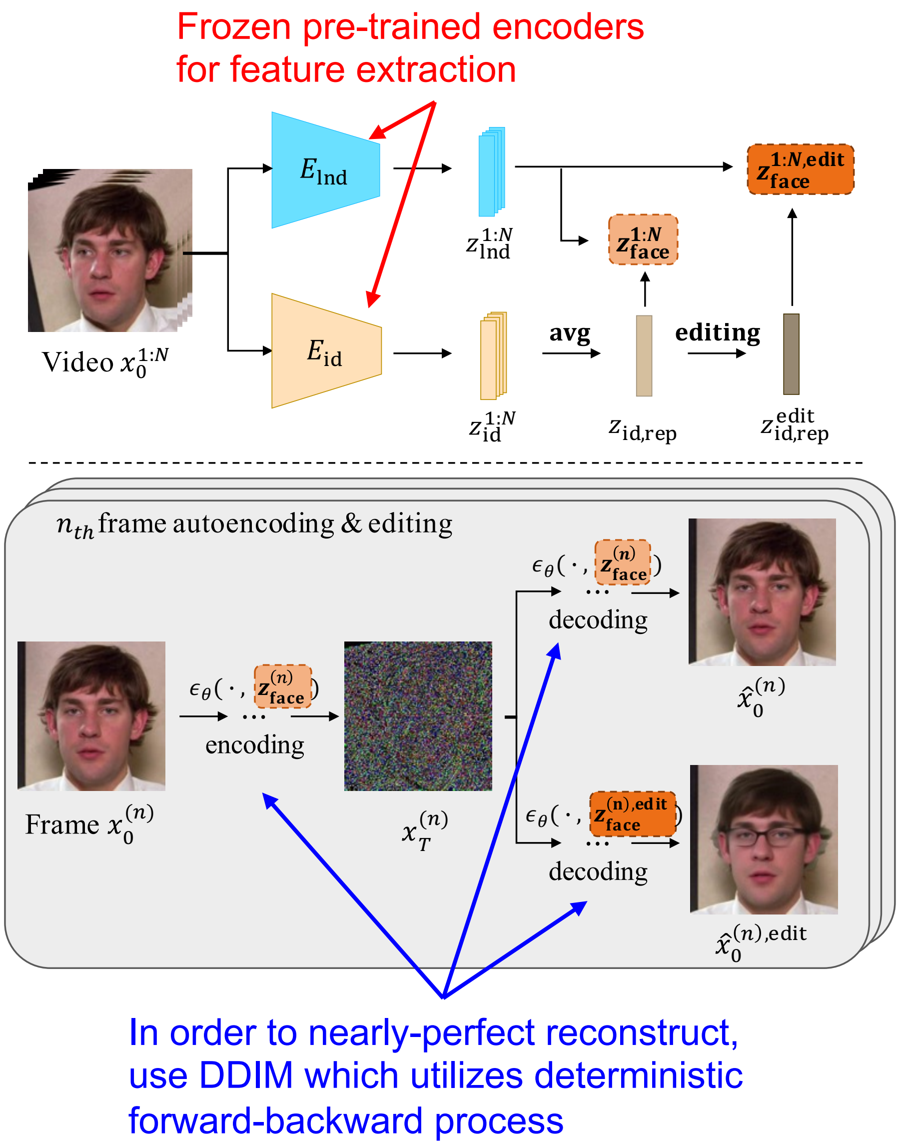
$$x_0^{(n)} \rightarrow \left(z_{\text{face}}^{(n)}, x_T^{(n)}\right) \rightarrow x_0^{(n)}$$

- High-level semantic latent $z_{\text{face}}^{(n)}$ (512-dim): consist of representative **identity** feature $z_{\text{id,rep}}$ and **motion** feature $z_{\text{lnd}}^{(n)}$

- Noise map $x_T^{(n)}$: Only information left out by $z_{\text{face}}^{(n)}$ is encoded (=**background** information)

- Since background information shows **high variance** to project to a low-dimensional space, encode background at noise map $x_T^{(n)}$

Method Overview: video autoencoding & editing pipeline



- Design a diffusion video autoencoder:

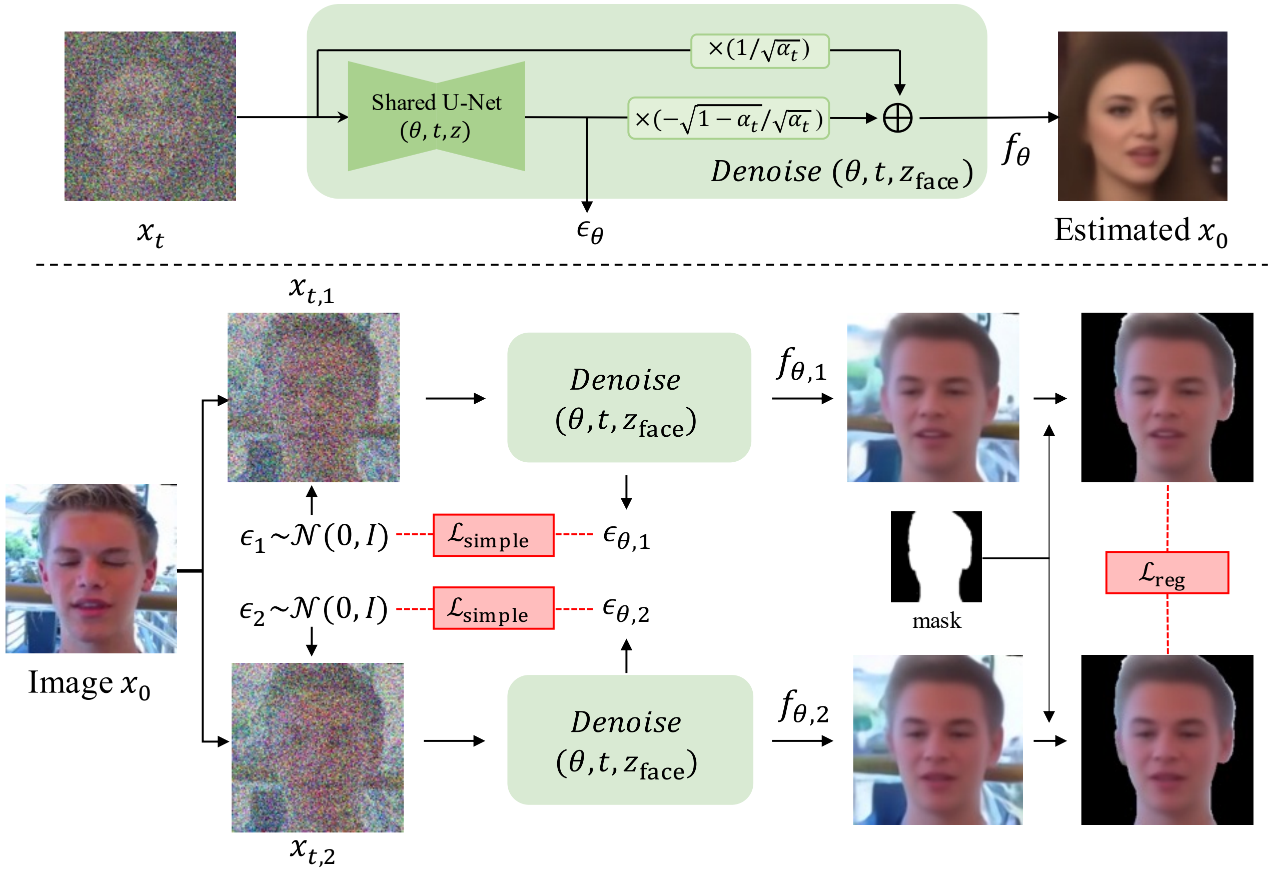
$$x_0^{(n)} \to \left(z_{\text{face}}^{(n)}, x_T^{(n)}\right) \to x_0^{(n)}$$

- High-level semantic latent $z_{\text{face}}^{(n)}$ (512-dim): consist of representative **identity** feature $z_{\text{id,rep}}$ and **motion** feature $z_{\text{lnd}}^{(n)}$

- Noise map $x_T^{(n)}$:
Only information left out by $z_{\text{face}}^{(n)}$ is encoded (=**background** information)

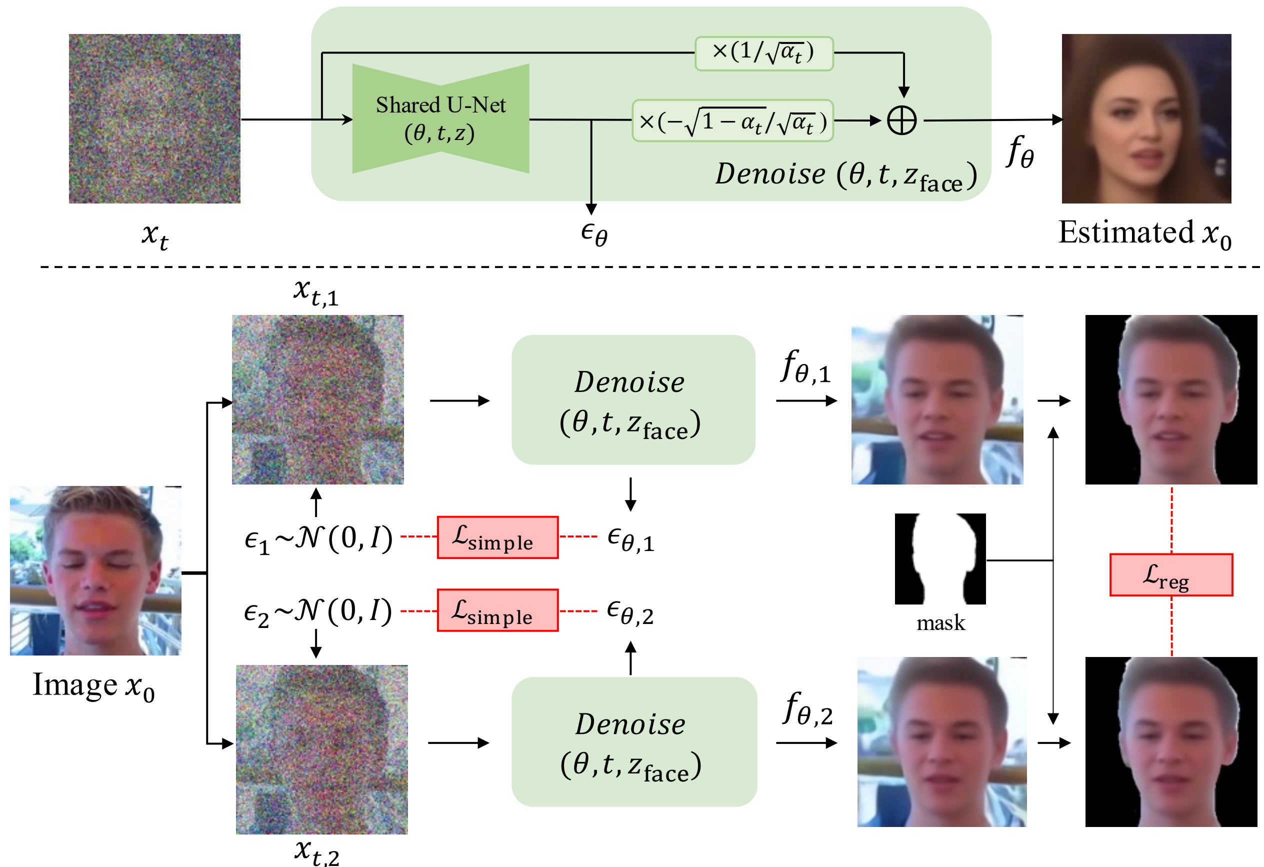
- Since background information shows **high variance** to project to a low-dimensional space, encode background at noise map $x_T^{(n)}$

Method Overview: training objective



- $\mathcal{L}_{simple} = \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(0,I), t} \|\epsilon_\theta(x_t, t, z_{face}) - \epsilon_t\|_1$
  - Simple version of DDPM loss
- $\mathcal{L}_{reg} = \mathbb{E}_{x_0 \sim q(x_0), \epsilon_1, \epsilon_2 \sim \mathcal{N}(0,I), t} \left\| f_{\theta,1} \odot m - f_{\theta,2} \odot m \right\|_1$
  - For clear decomposition btw background and face information

Method Overview: training objective



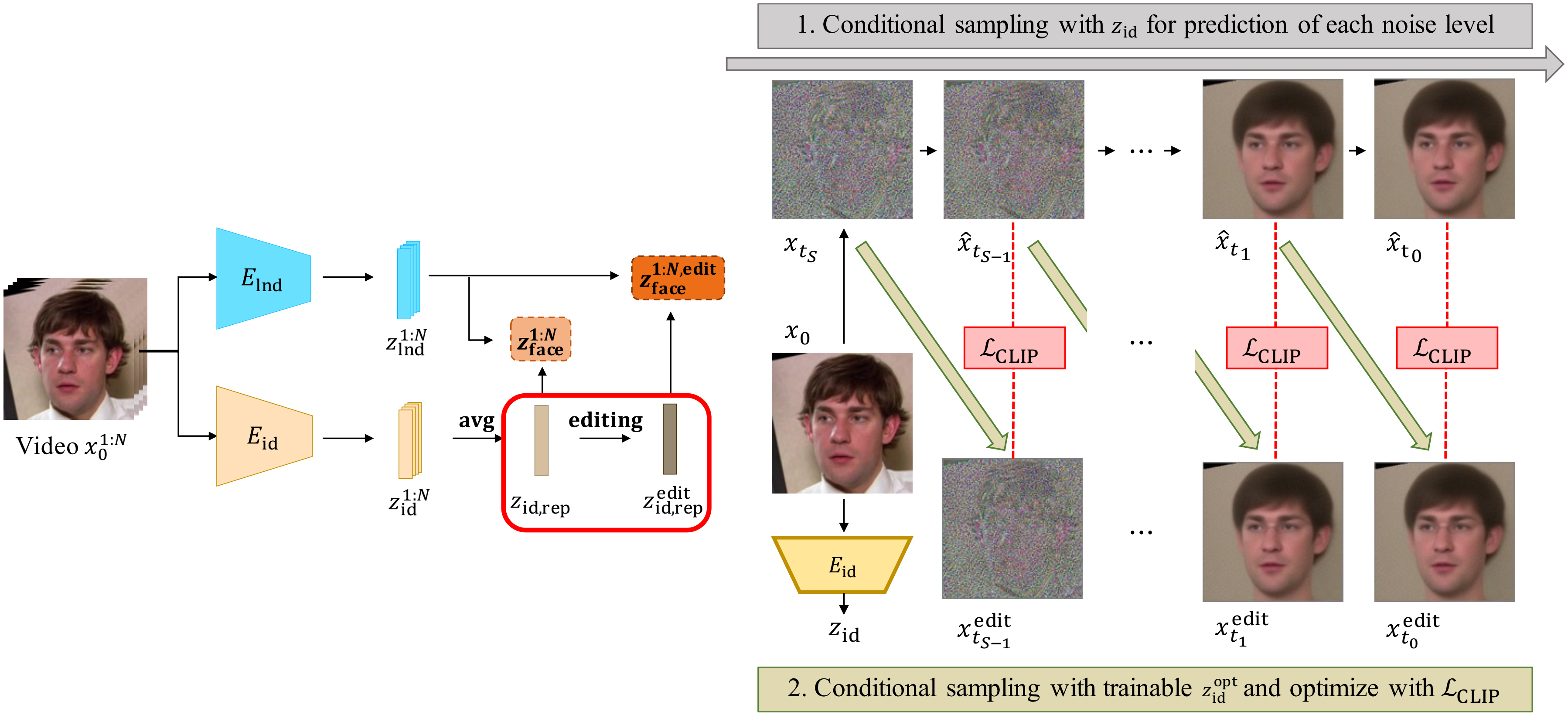
- $\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(0,I), t} \| \epsilon_\theta(x_t, t, z_{\text{face}}) - \epsilon_t \|_1$
  - Simple version of DDPM loss

Encourages the useful information of the image to be well contained in the semantic latent z_{face}

- $\mathcal{L}_{\text{reg}} = \mathbb{E}_{x_0 \sim q(x_0), \epsilon_1, \epsilon_2 \sim \mathcal{N}(0,I), t} \left\| f_{\theta,1} \odot m - f_{\theta,2} \odot m \right\|_1$
  - For clear decomposition btw background and face information

Effect of noise in x_t on the face region will be reduced and z_{face} will be responsible for face features

Method Overview: video editing framework



- Classifier-based editing
  - Train a linear classifier for each attribute of CelebA-HQ in the identity feature z_{id} space
- CLIP-based editing
  - Minimize CLIP loss between intermediate images with drastically reduced number of steps S ($\ll T$)

Experiment: Reconstruction

Table 1. **Quantitative reconstruction results** on the randomly chosen 20 videos in VoxCeleb1 test set. The reported values are the mean of the averaged per-frame measurements for each video.

Method	SSIM ↑	MS-SSIM ↑	LPIPS ↓	MSE ↓	
e4e [34]	0.509	0.761	0.157	0.037	← Latent Transformer
PTI [27]	0.765	0.939	0.063	0.007	← STIT
Ours ($T = 20$)	0.540	0.905	0.228	0.016	
Ours ($T = 100$)	**0.922**	**0.989**	**0.045**	**0.002**	

- Our diffusion video autoencoder with T = 100 shows the **best reconstruction ability** and still outperforms e4e with only T = 20

Experiment: Temporal Consistency

Table 2. **Quantitative results** to evaluate temporal consistency. Ours show the best global coherency and comparable local consistency to the baselines.

Method	TL-ID	TG-ID
Yao *et al.* [41]	0.989	0.920
Tzaban *et al.* [35]	0.997	0.961
Xu *et al.* [40]	1.002	0.983
Ours	0.995	0.996

← interpret as being consistent as the original is when their values are close to 1

Original | Latent Transformer | STIT | VideoEditGAN | Ours

- Only our diffusion video autoencoder successfully produces the **temporally consistent** result
- We greatly improve global consistency (TG-ID)

Experiment: Editing Wild Face Videos

"young"

"gender"

Original | Latent Transformer | STIT | Ours

"beard"

Original | STIT | Ours

- Owing to the reconstructability of diffusion models, editing **wild videos** that are difficult to inversion by GAN-based methods becomes possible.

Experiment: Decomposed Features Analysis

Input | Random x_T | Identity switch | Motion switch | Background switch

- Generated images with switched identity, motion, and background feature **confirm** that the features are **properly decomposed**

Experiment: Ablation Study

- Without the regularization loss, the **identity changes significantly** according to the **random noise**
  - we can conclude that the regularization loss helps the model to decompose features effectively

MLILAB Machine Learning & Intelligence

Conclusions

- Our contribution is four-fold:
  - We **devise** diffusion video autoencoders that decompose the video into a single time-invariant and per-frame time-variant features for temporally consistent editing
  - Based on the decomposed representation of our model, face video editing can be **conducted** by editing only the single time-invariant identity feature and decoding it together with the remaining original features
  - Owing to the nearly-perfect reconstruction ability of diffusion models, our framework can be utilized to edit **exceptional cases** such that a face is partially occluded by some objects as well as usual cases
  - In addition to the existing predefined attributes editing method, we propose a text-based identity editing method based on the local directional CLIP loss for the **intermediately generated product** of diffusion video autoencoders

Thank you !

Any Questions ?