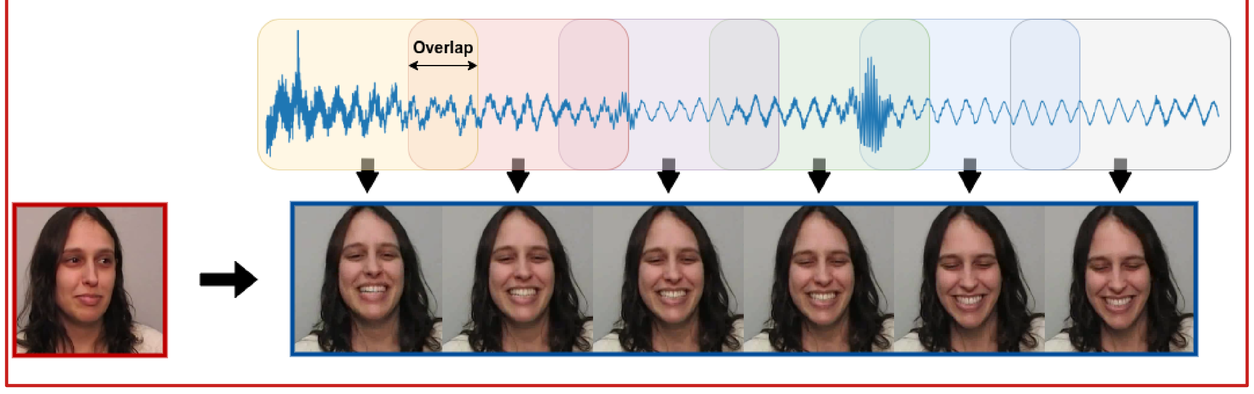
Laughing Matters: Introducing Audio-Driven Laughing-Face Generation with Diffusion Models

Antoni Bigata Casademunt Rodrigo Mira Nikita Drobyshev Konstantinos Vougioukas Stavros Petridis Maja Pantic
ab4522@ic.ac.uk rs2517@ic.ac.uk nikita.drobyshev23@gmail.com k.vougioukas@ic.ac.uk stavros.petridis04@ic.ac.uk
m.pantic@ic.ac.uk
Intelligent Behaviour Understanding Group (iBUG)
Imperial College London

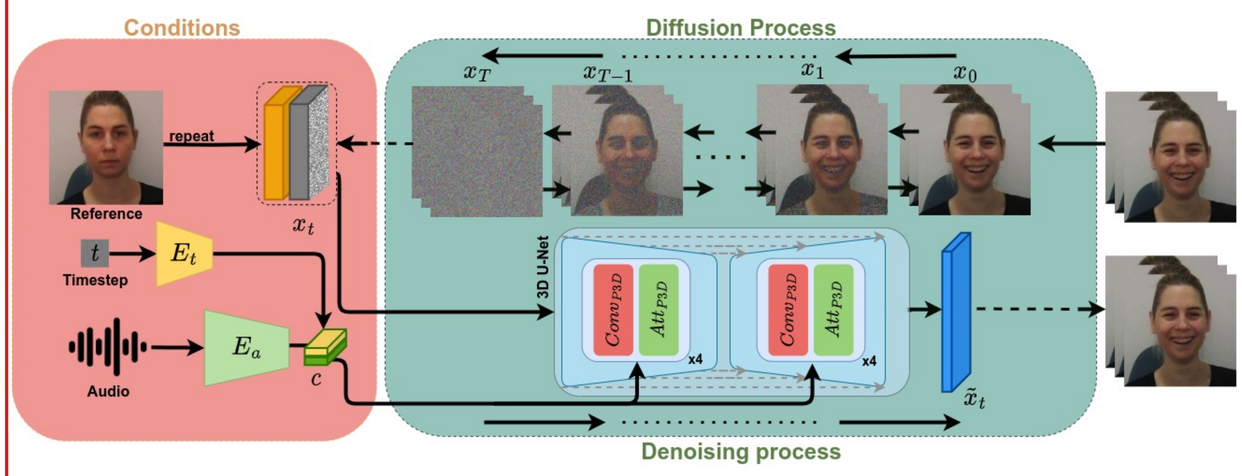
BMVC 2023

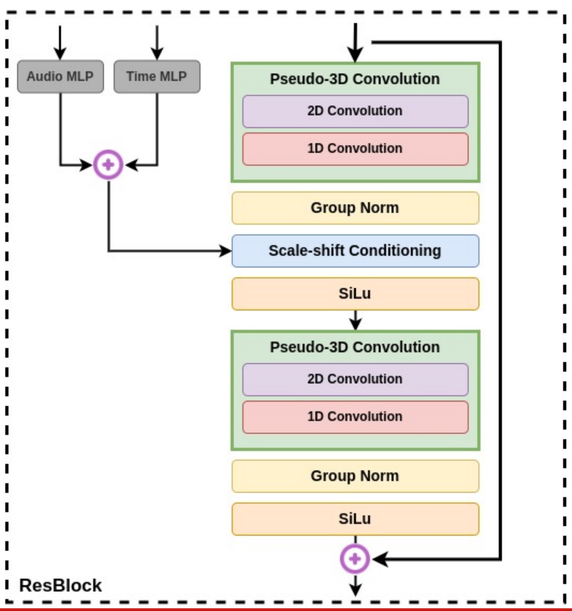
Introduction

- Facial animation is crucial for immersive experiences in VR, movies, and HCI. However, current methods focus on speech-driven animation, neglecting non-verbal expressions like laughter, which convey crucial context.
- Laughter's complexity and lack of direct audio-visual correlation make it challenging to generate realistically. Recent advances in speech-driven animation show promise, but laughter remains a unique challenge.
- A novel video diffusion model is introduced in this paper to generate natural laughter videos, addressing the shortcomings of existing methods and outperforming them in producing synchronized and realistic laughter animations.



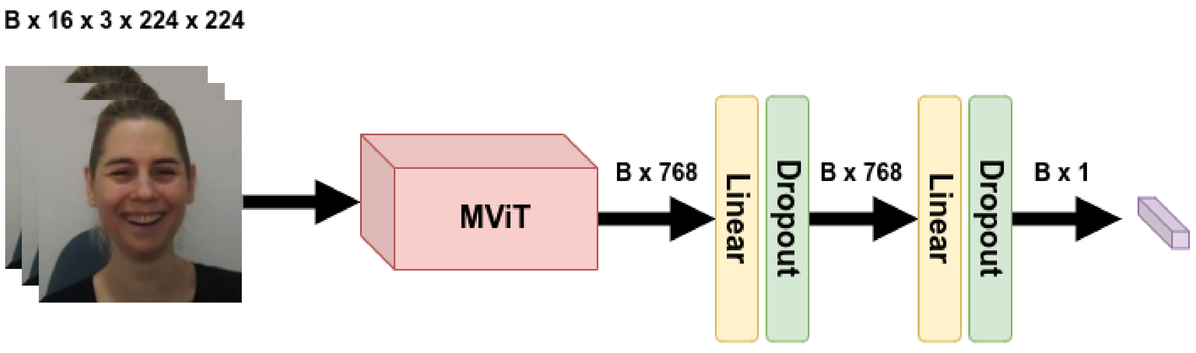
Approach





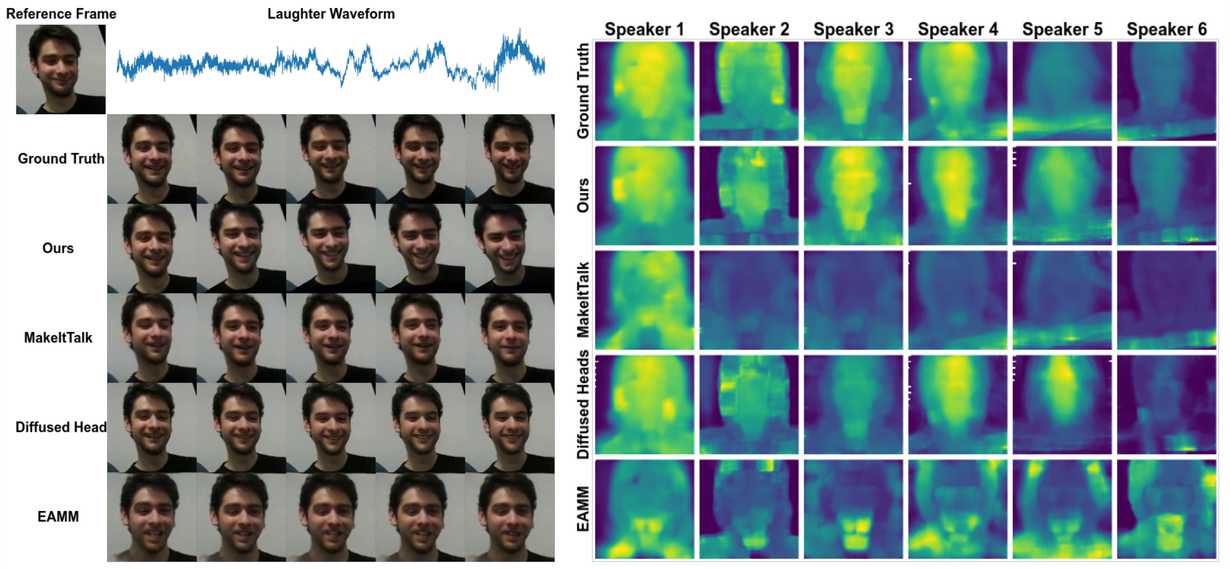
- Condition signal includes a single frame and audio sequence processed by an audio encoder.
- Pseudo-3D Convolutions and Attention Layers to balance computational efficiency and information sharing.
- Conditioning is applied in each ResBlock using a scale-shift operation with scale and shift derived from audio and timestep information.

Evaluation Metrics



- Evaluation metrics include FID, SSIM, and FVD for image and video quality assessment.
- A Laughter Classifier (LC) is trained to distinguish between speech and laughter videos using a pre-trained MViTv2 model fine-tuned with MAHNOB data.
- The LC is used to showcases the limitations of pre-trained speech animation models in laughter generation

Qualitative Evaluation



- Compared to competing approaches like EAMM and MakeItTalk, our model preserves identity and synchronizes head movement with audio input.
- Optical flow analysis demonstrates that our model closely matches the ground truth for natural movement, outperforming MakeItTalk, EAMM, and Diffused Heads, especially for certain speakers (3, 5, and 6).

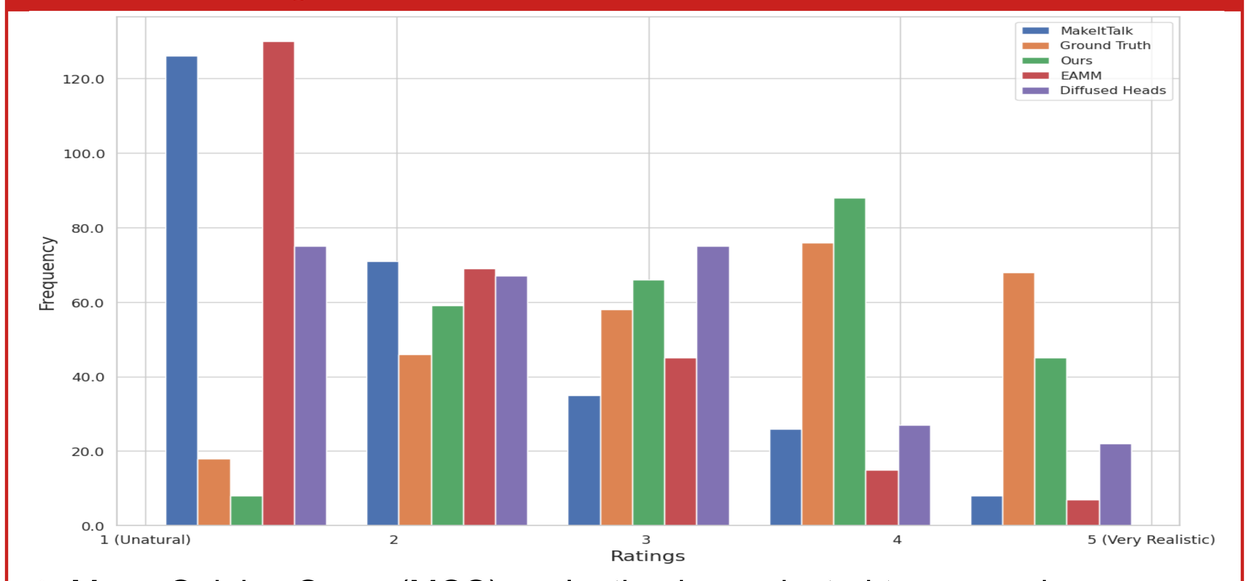
Comparison with other works

Model	FVD ↓	FID ↓	SSIM ↑	LC ↑ (%)	MOS ↑
Pre-trained					
Diffused Heads [50]	149.51	49.36	0.236	80.70	-
SDA [55]	594.32	111.89	0.053	13.85	-
EAMM [23]	391.62	71.71	0.094	16.67	-
PC-AVS [61]	1164.49	175.99	0.004	53.91	-
MakeItTalk [62]	196.89	49.08	0.262	72.50	1.94±1.12
Re-trained					
Diffused Heads [50]	152.30	67.46	0.232	94.09	2.45±1.22
SDA [55]	696.33	124.52	0.040	85.13	-
EAMM [23]	324.97	74.18	0.095	20.67	1.87±1.05
Laughing Matters (Ours)	**111.95**	**45.69**	**0.371**	**96.52**	**3.39±1.09**
Ground truth	-	-	-	100.00	3.49±1.23

Ablation study

Training configuration	FVD ↓	FID ↓	SSIM ↑	LC ↑ (%)
Baseline	**111.95**	**45.69**	**0.371**	**96.52**
w/o Augmentation regularization	195.03	60.60	0.308	83.93
w/o Classifier-free guidance	126.89	46.91	0.302	75.09

User study



- Mean Opinion Score (MOS) evaluation is conducted to assess human perception of video quality, accounting for the limitations of quantitative metrics.

References

[23] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. EAMM: one-shot emotional talking face via audio-based emotion-aware motion model.
[50] Michal Stypulkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation.
[55] onstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans.
[61] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation.
[62] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation.

Supplementary videos and code available at: