

EMO: Emote Portrait Alive

汇报人：项一卓

2024.06



Reference Image



Generated
Video

EMO

标题: EMO: Emote Portrait Alive – Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions

作者: Linrui Tian, Qi Wang, Bang Zhang, Liefeng Bo

单位: Institute for Intelligent Computing, Alibaba Group

项目地址: <https://humanaigc.github.io/emote-portrait-alive/>

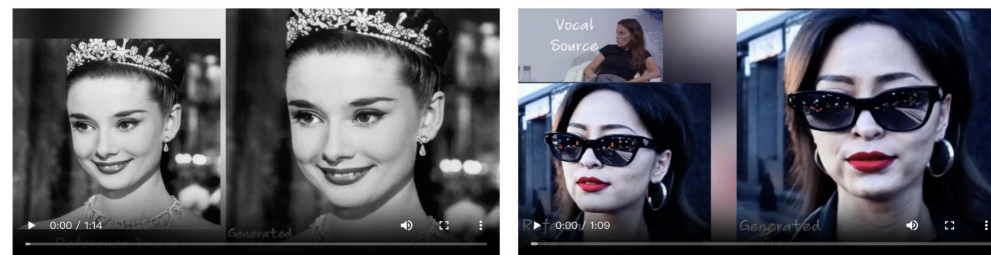
GitHub: <https://github.com/HumanAIGC/EMO> (7.5k)

被引: 13

EMO: Emote Portrait Alive - Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions

Linrui Tian, Qi Wang, Bang Zhang, Liefeng Bo
Institute for Intelligent Computing, Alibaba Group

GitHub arXiv



Character: Audrey Kathleen Hepburn-Ruston

Character: Al Lady from SORA

Google Scholar



Liefeng Bo

Head of Applied Computer Vision Lab at Alibaba Group
Verified email at cs.washington.edu - [Homepage](#)
Machine Learning Computer Vision Robotics

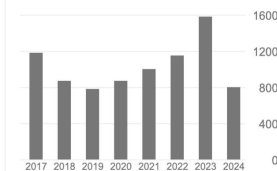
FOLLOW

GET MY OWN PROFILE

TITLE	CITED BY	YEAR
A large-scale hierarchical multi-view rgb-d object dataset K Lai, L Bo, X Ren, D Fox Robotics and Automation (ICRA), 2011 IEEE International Conference on, 1817-1824	1807	2011
Multiobjective immune algorithm with nondominated neighbor-based selection M Gong, L Jiao, H Du, L Bo Evolutionary computation 16 (2), 225-255	615	2008
RGB-(D) Scene Labeling: Features and Algorithms X Ren, L Bo, D Fox International Conference on Computer Vision and Pattern Recognition (CVPR)	582	2012
Unsupervised Feature Learning for RGB-D Based Object Recognition L Bo, X Ren, D Fox International Symposium on Experimental Robotics (ISER)	503	2012
Kernel descriptors for visual recognition L Bo, X Ren, D Fox Advances in Neural Information Processing Systems (NIPS)	460	2010
Unsupervised Feature Learning for 3D Scene Labeling K Lai, L Bo, D Fox Robotics and Automation (ICRA), 2014 IEEE International Conference on	414	2014

Cited by [VIEW ALL](#)

	All	Since 2019
Citations	14119	6219
h-index	54	42
i10-index	96	77



Public access [VIEW ALL](#)

Public access	Count
0 articles	21 articles
not available	available

Based on funding mandates

目录

CONTENT

01 一、EMO模型简介

02 二、模型的输入与输出

03 三、模型的工作原理

04 四、训练过程

05 五、创新与应用

06 六、结论



01

模型简介

EMO – 网络结构

Backbone: SD 1.5 (Cross-Attn to Ref-Attn)

Audio: $A_{gen}^{(f)} = \oplus \{A^{(f-m)}, \dots, A^{(f)}, \dots, A^{(f+m)}\}$

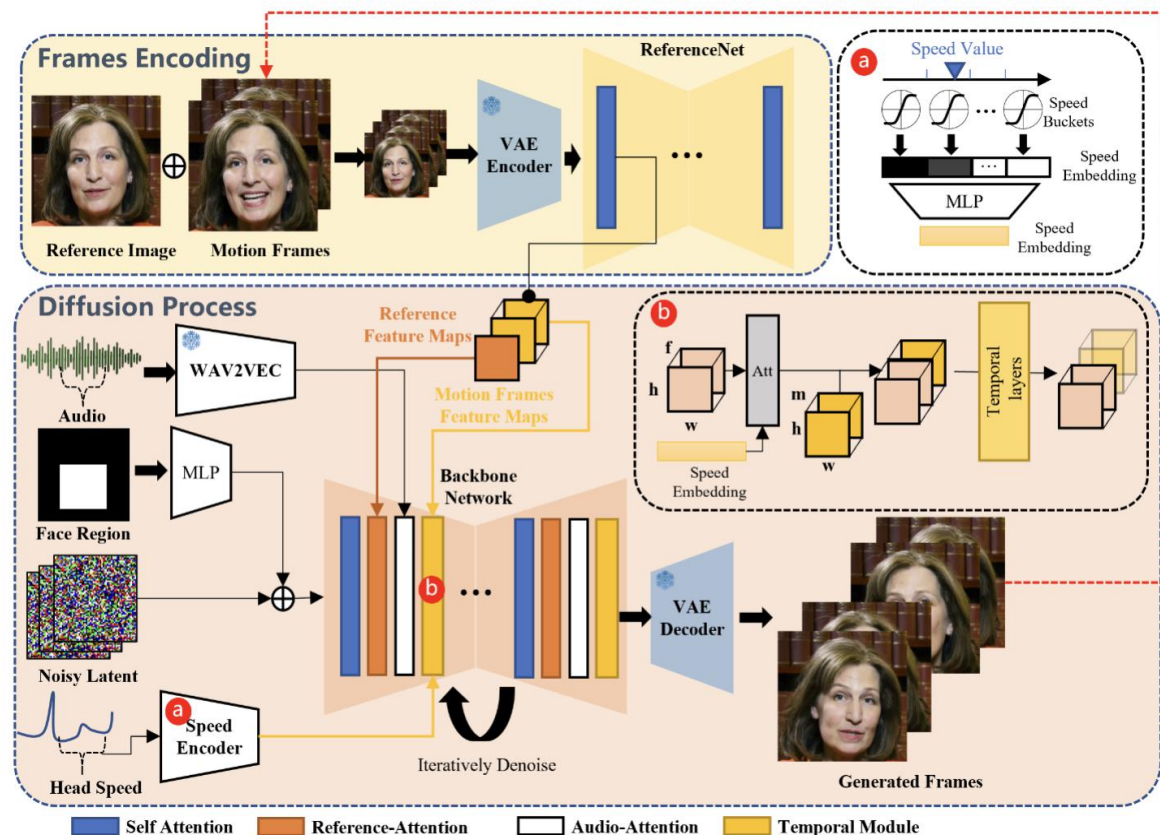
ReferenceNet: SD 1.5


Temporal: AnimateDiff (add Motion Frames Feature Maps)

Face Locator: Conv Layers

Speed Encoder: MLP + Cross-Attn

To address this issue, we incorporate the target head motion speed into the generation. More precisely, we consider the head rotation velocity w^f in frame f and divide the range of speeds into d discrete speed buckets, each representing a different velocity level. Each bucket has a central value c^d and a radius r^d . We retarget w^f to a vector $S = \{s^d\} \in \mathbb{R}^d$, where $s^d = \tanh((w^f - c^d)/r^d * 3)$. Similar to the method used in the audio layers, the head rotation speed embedding for each frame is given by $S^f = \oplus \{S^{(f-m)}, \dots, S^{(f)}, \dots, S^{(f+m)}\}$, and $S^f \in \mathbb{R}^{b \times f \times c^{speed}}$ is subsequently processed by an MLP to extract speed features.

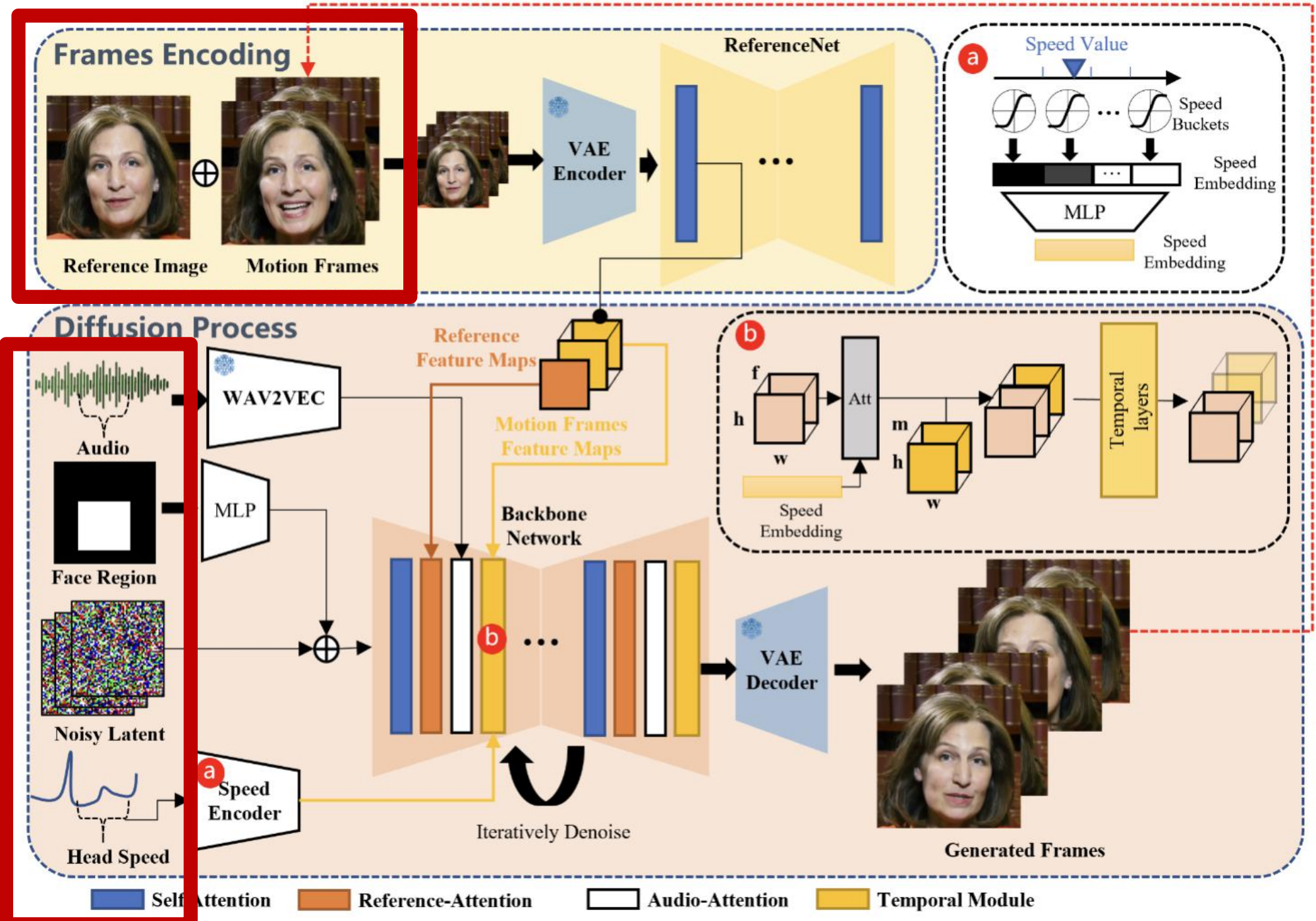





02

输入与输出

- 图像
- 音频
- Face Region
- Head Speed





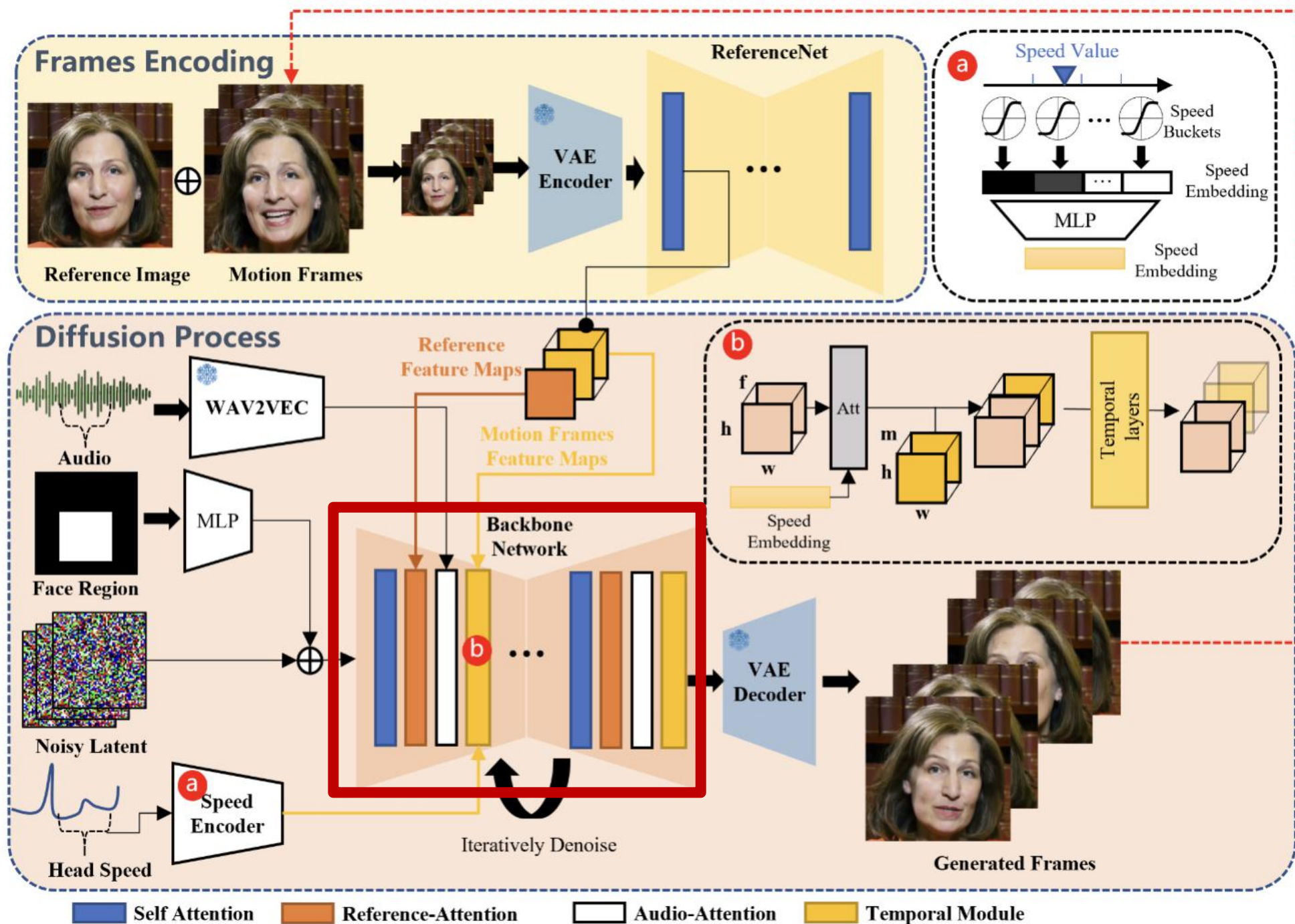
03

工作原理

Backbone

1、特征对齐与生成 – 多种Attention机制 (红+橙+白)

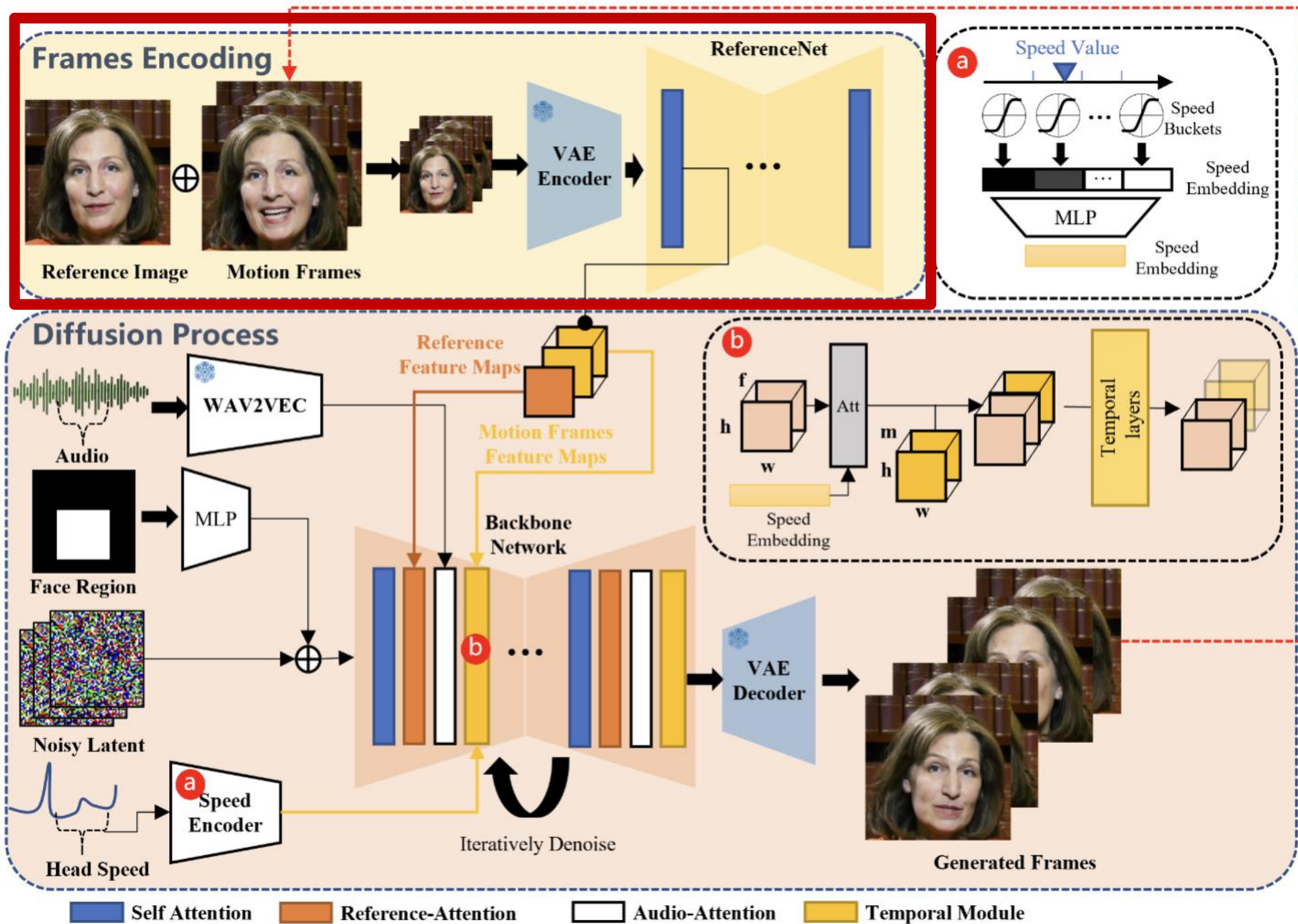
- Self-Attention (Stable Diffusion 1.5)
- Reference-Attention
- Audio-Attention
- Temporal Module



Backbone

2、细看 Reference

- ReferenceNet
- motion frames

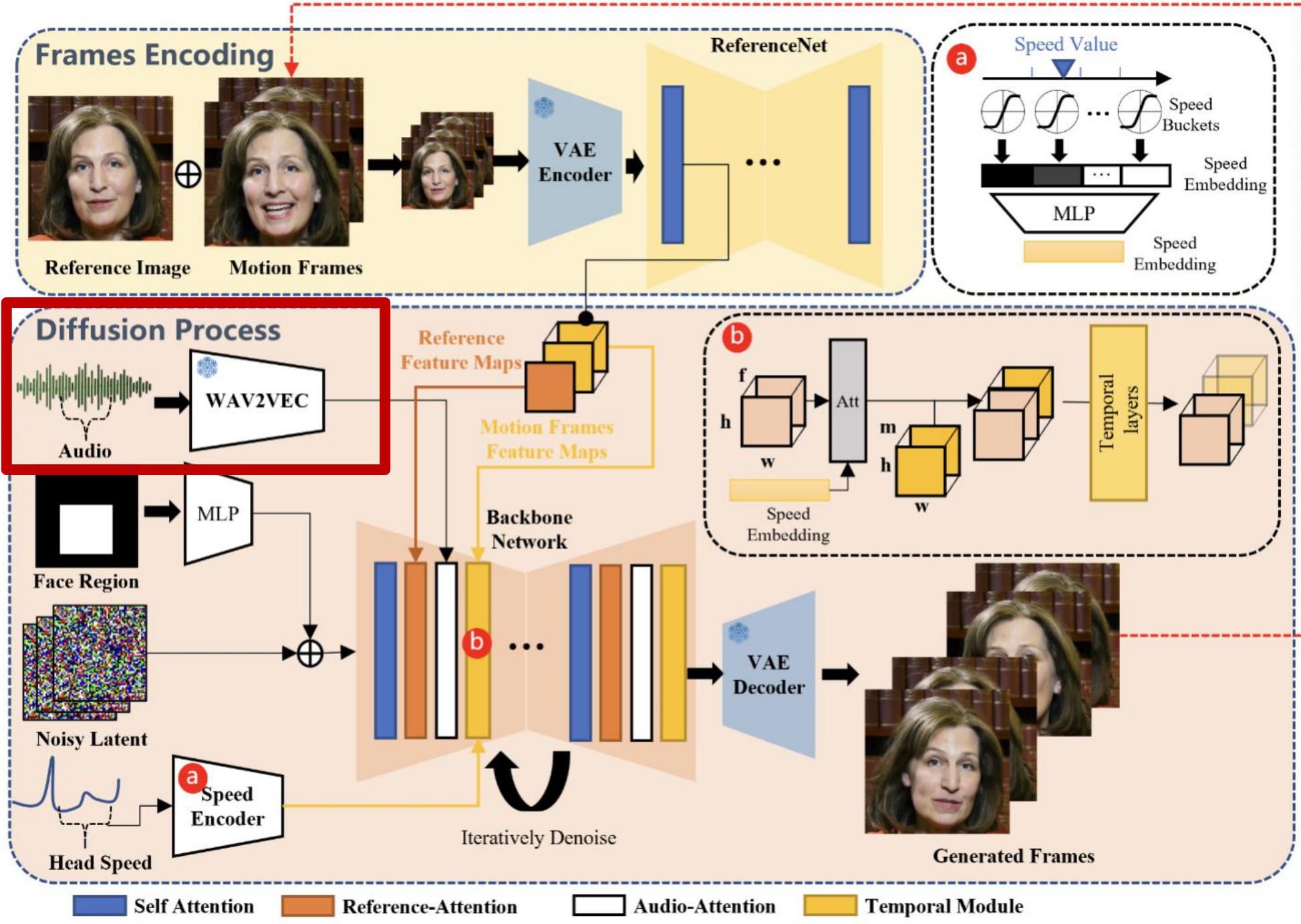


Backbone

3、Audio (白)

• V2V

(Voice to Vector)



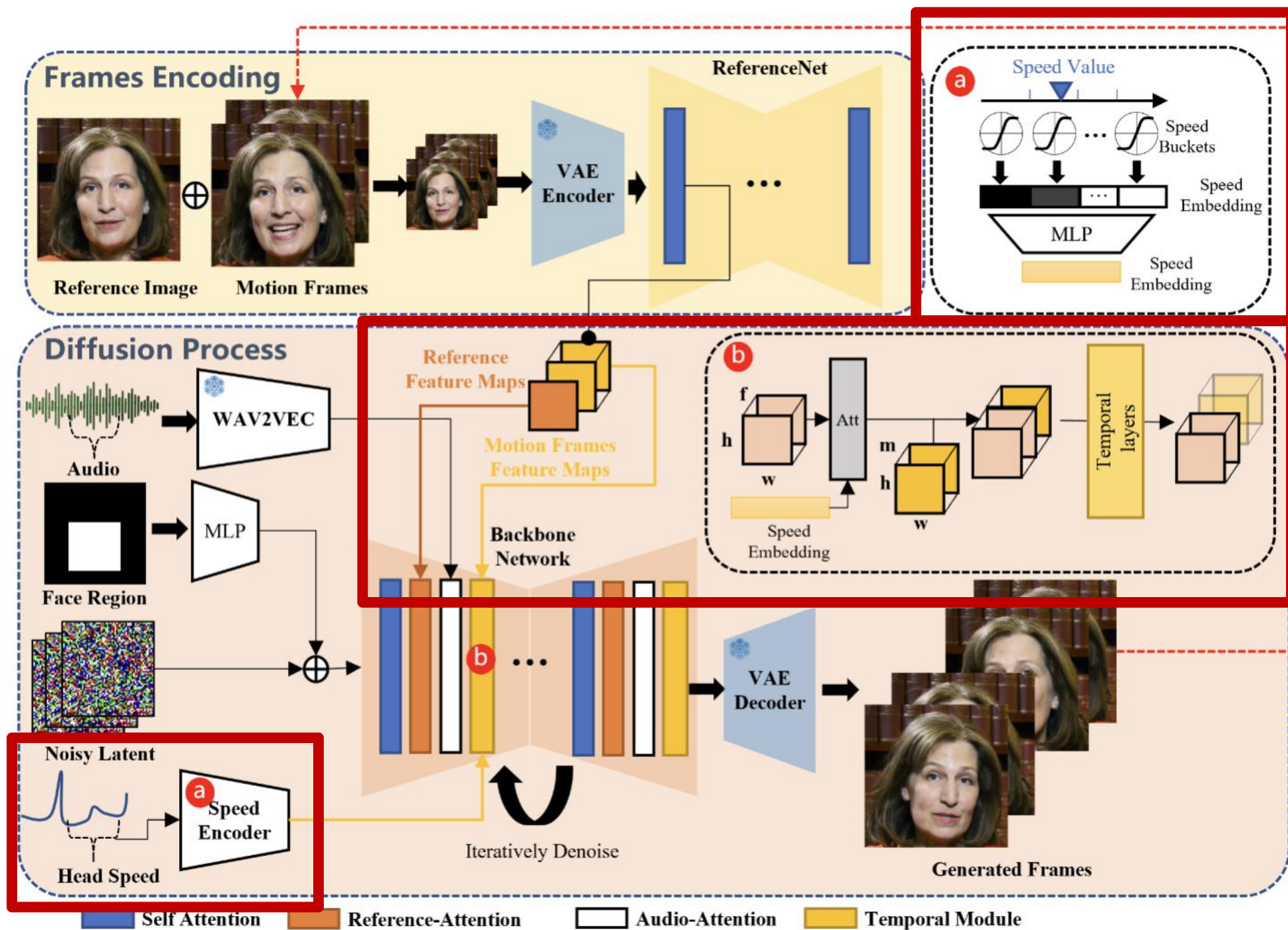
Backbone

4、连贯视频生成(黄)

• Temporal Module

• Animate Diffusion

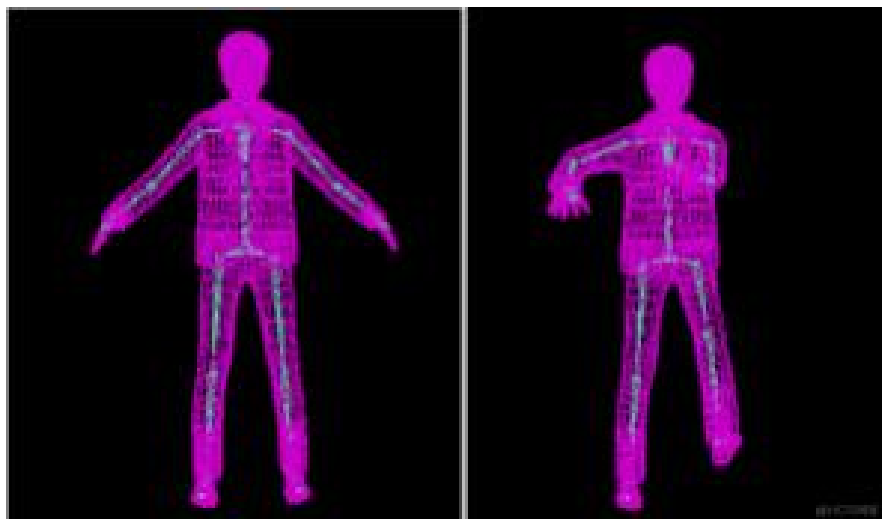
• Speed Encoder
(bucketing (分桶),
embedding)



深度学习中弱控制的优势与弊端：

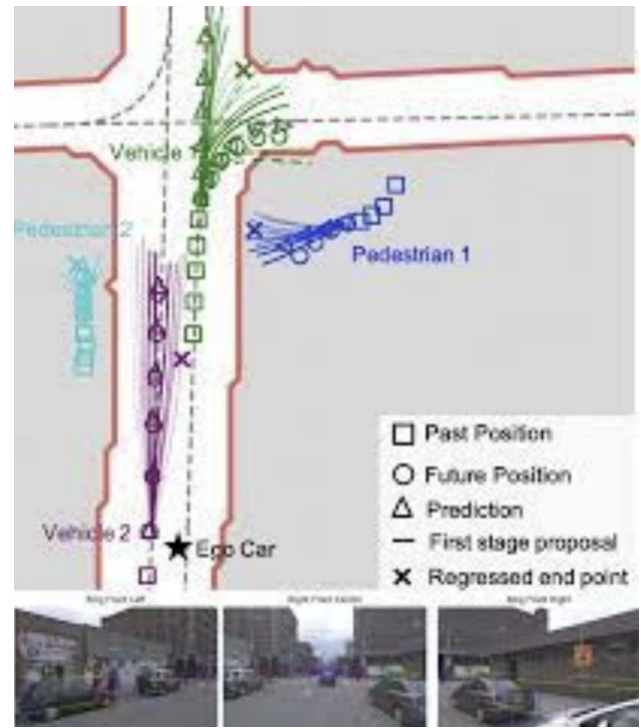
优势：灵活性高、训练速度快

弊端：精度有限



稠密控制：

骨架标注、面部标注、语义分割、轨迹标注





04

训练过程

Stage 1. Image pretraining: **Backbone Network** takes a single frame as input. **ReferenceNet** handles a distinct randomly chosen frame from the same video clip, **Face Locator**

Stage 2. Video training: **temporal modules** and **audio layers**, $n+f$ contiguous frames are sampled from the video clip

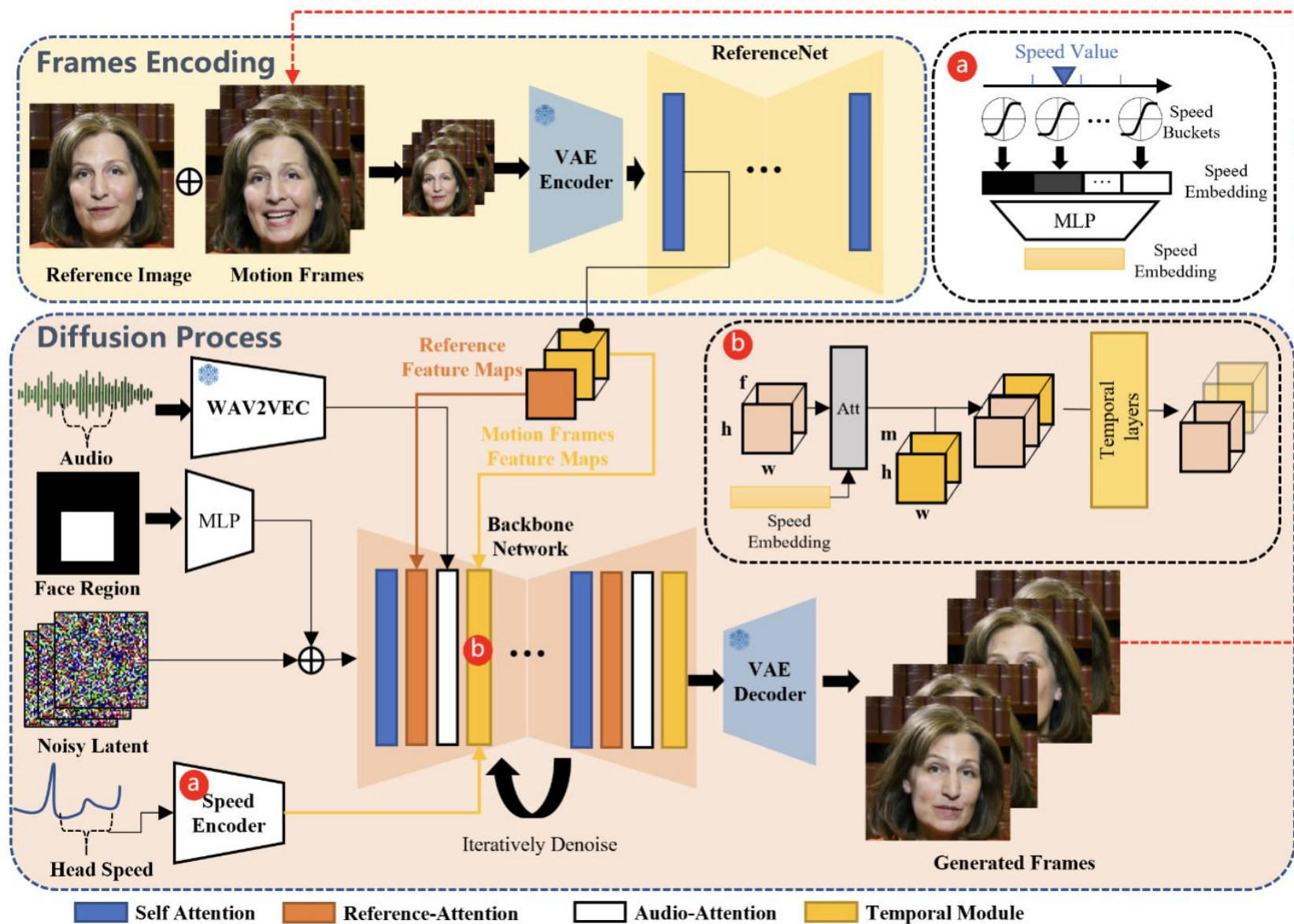
Stage 3. **speed layers** and **temporal modules** (simultaneous training of both the speed and audio layers undermines the driven ability of the audio on the character's motions.)

We collected approximately **250 hours of talking head videos** from the internet and supplemented this with the HDTF and VFHQ datasets to train our models.

Head rotation velocity was labeled by extracting the 6-DoF head pose for each frame using facial landmarks, followed by calculating the rotational degrees between successive frames.

训练过程

1. 纯图像生成训练
2. 视频生成训练
3. 头部速度控制训练



谢谢大家!

汇报人：项一卓

2024.06