Full-body Animation Generation for Expressive NPCs

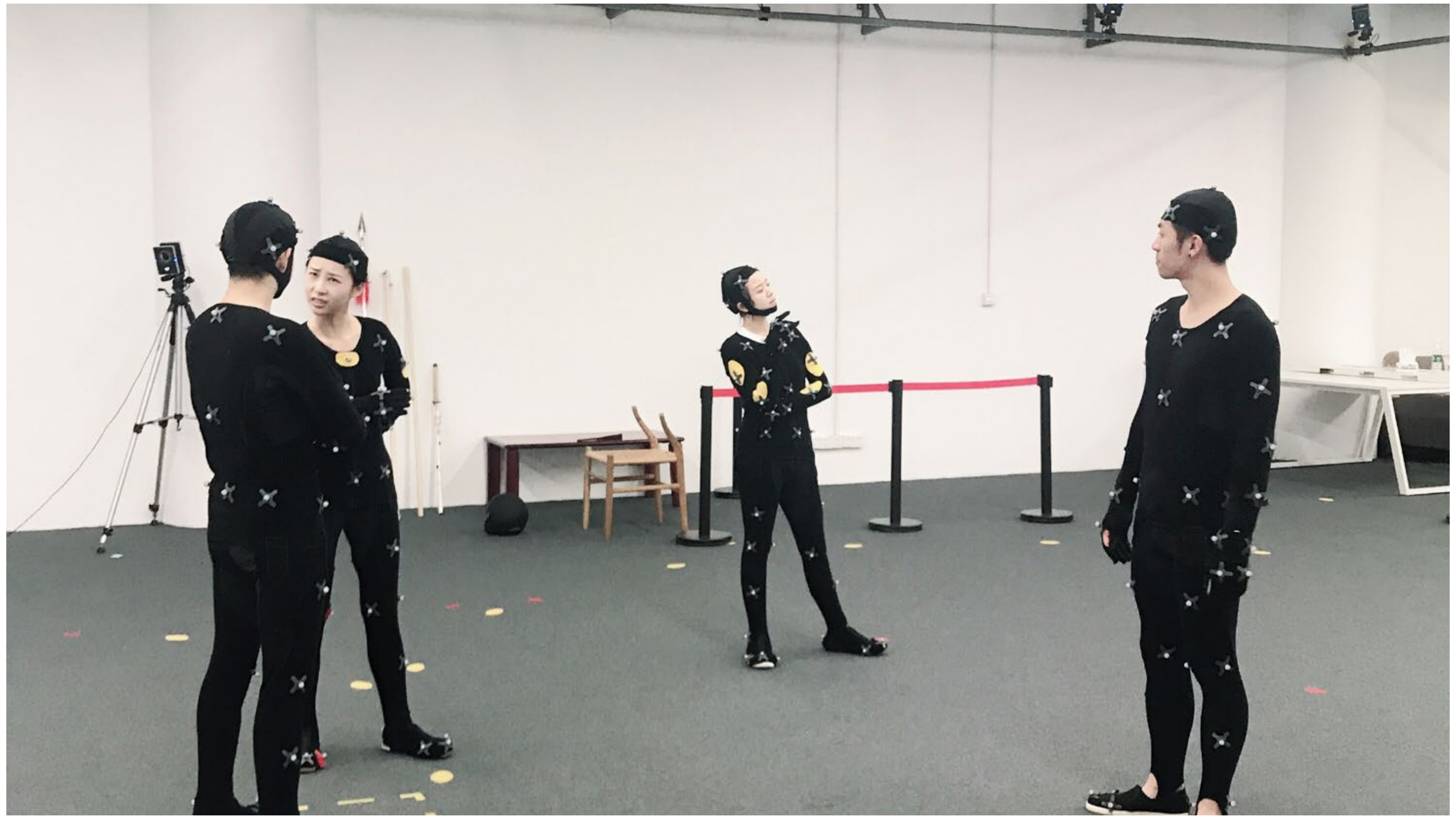
Yu Ding
Artificial Intelligence Expert & Netease Fuxi AI Lab, China

Problem Definition

Existing Methods and Limitations

Handcraft & Motion capture:

- Labor intensive
- Manual correction
- Time-consuming
- Planned scenarios
- Expensive



Concatenating Animation Clips:

- Repetitive
- Interpolation
- Missing subtle animations

Question:

How to automatically generate full-body animations?

Objective

Automatic generation of the full-body animations for talking NPCs

- High-efficiency: online and real-time
  - ✓ Face expression < 60ms for a 15-second sequence
  - ✓ Body language < 600ms for a 15-second sequence
- High-quality:
  - ✓ natural, lifelike, human-like
  - ✓ expressive (emotion and intention)
  - ✓ variety

Jusice

A Chinese Ghost Story

SONG DYNASTY CINEMA

Revelation Mobile

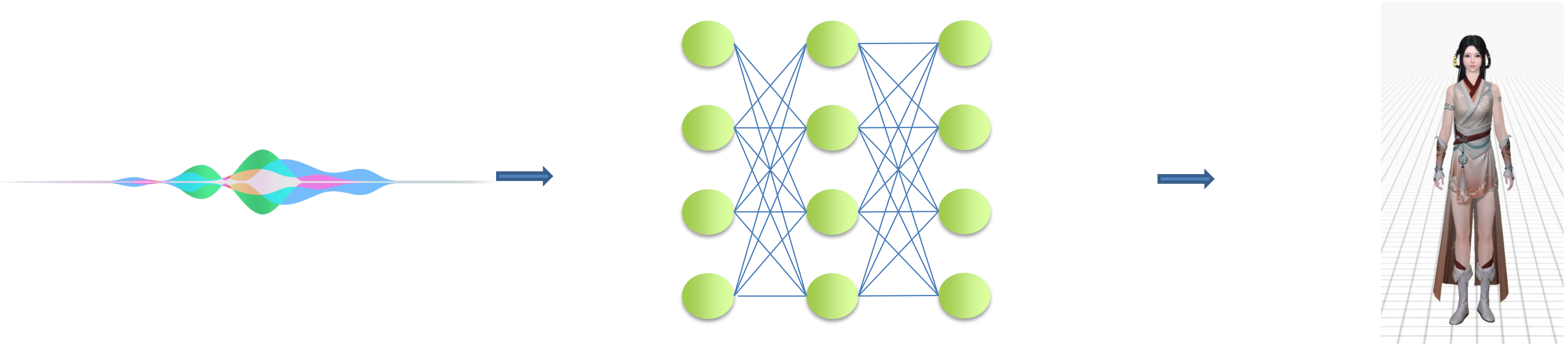
Background

While speaking, humans make behaviors to

- Emphasize ideas
- Signal syntactic boundaries and stress
- Complement verbal information
- Express emotions/intention

✓ Human behaviors are closely related to simultaneous speech.

Solution: speech-to-animation



Input Speech

Neural Networks

Full-body Animations

Definition: full-body animation

- 28D lower facial expression: lip and jaw
- 22D upper facial expression: eyelids and eyebrows
- 76D body gestures: head, hand, torso, and legs

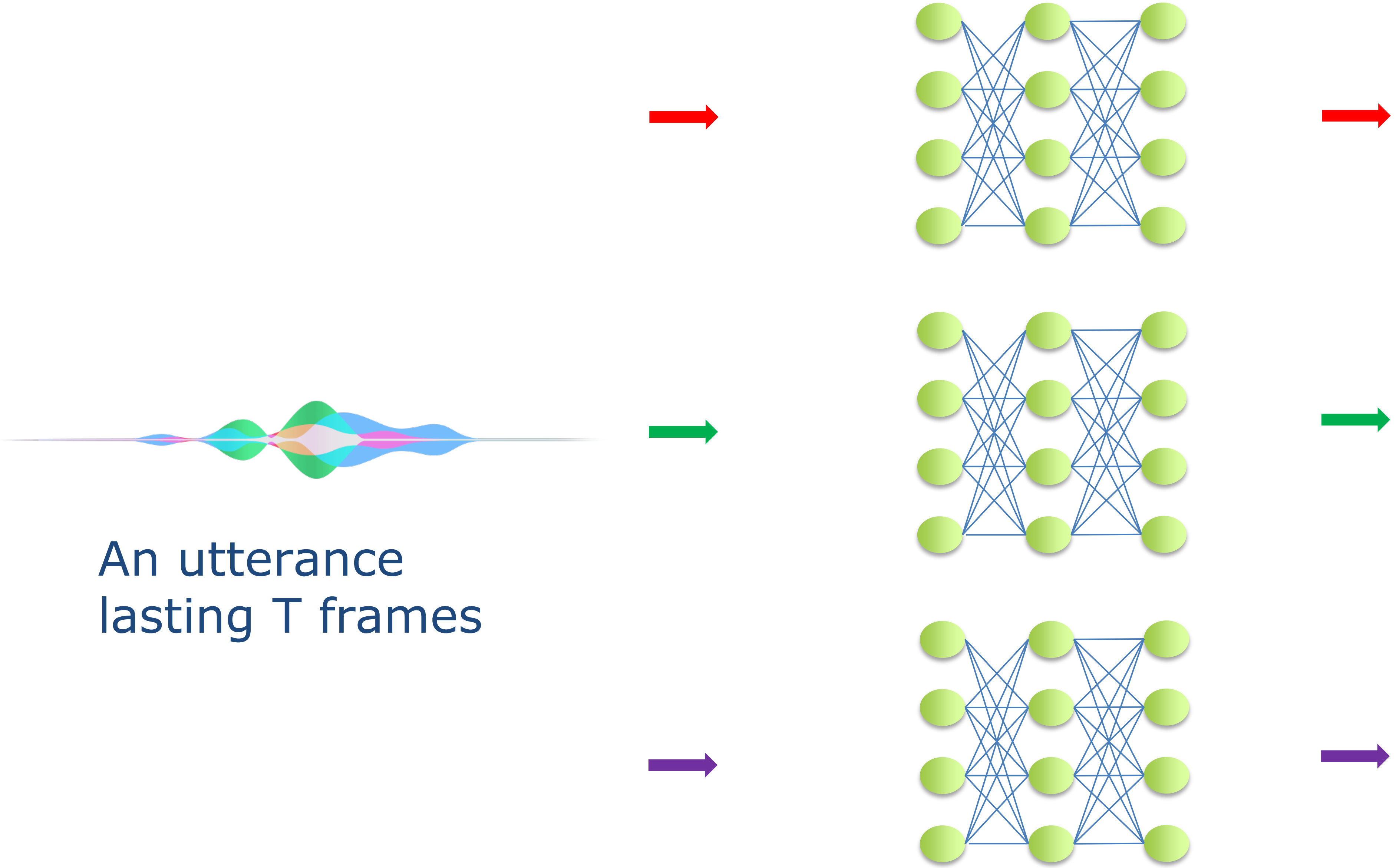
Requiring high-quality motion capture data and manual retargeting

Methodology

Input Speech

Three Neural Networks

Full-body Animations



An utterance lasting T frames

- Lower facial expression

  T X 28 animation sequence

- Upper facial expression

  T X 23 animation sequence

- Body gestures

  T X 76 animation sequence

Neural Network: Lower Facial Expression

- Lips
- Jaw

Neural Network: Lower Facial Expression

Trick1: using time-aligned phonemes, instead of speech features (e.g. MFCC)
Why: speech features entangled from speaker timbre non-related to animation

Synthetic Speech -> Time aligned phonemes
Human speech -> ASR-> Time aligned phonemes

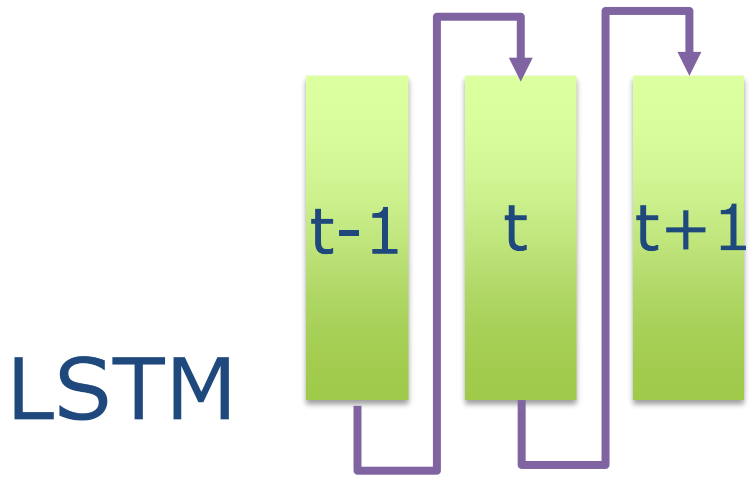
Spoken sentence: oh yeah, you don't want to tell me where she is.

Audio

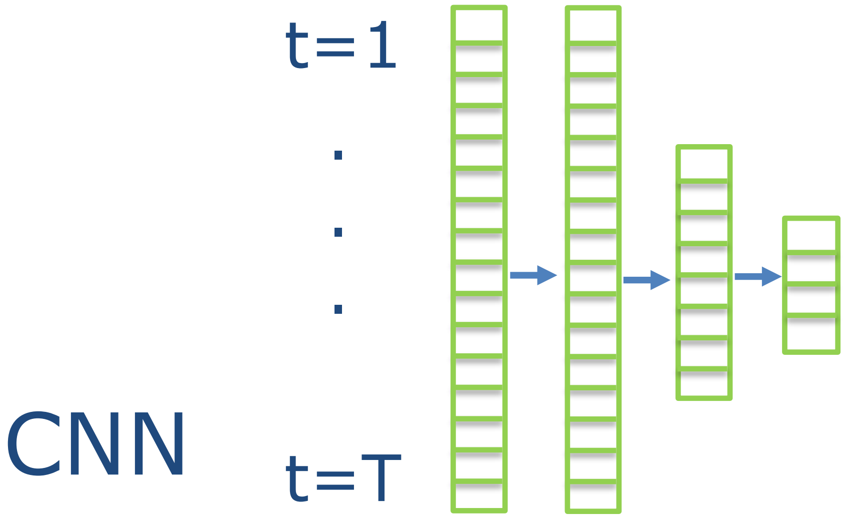
Time aligned phonemes: $ph = \{ow, ow, ow, \cdots, sh, sh, sh, ih, ih, ih, z, z, z\}$

Neural Network: Lower Facial Expression

Trick2: using CNN-based model, instead of LSTM-based ones
Why: CNN-based model is also capable of processing sequential problems,
but also parallel computing. (CNN 20-60ms v.s. LSTM 500-800ms for a 15-second clip)

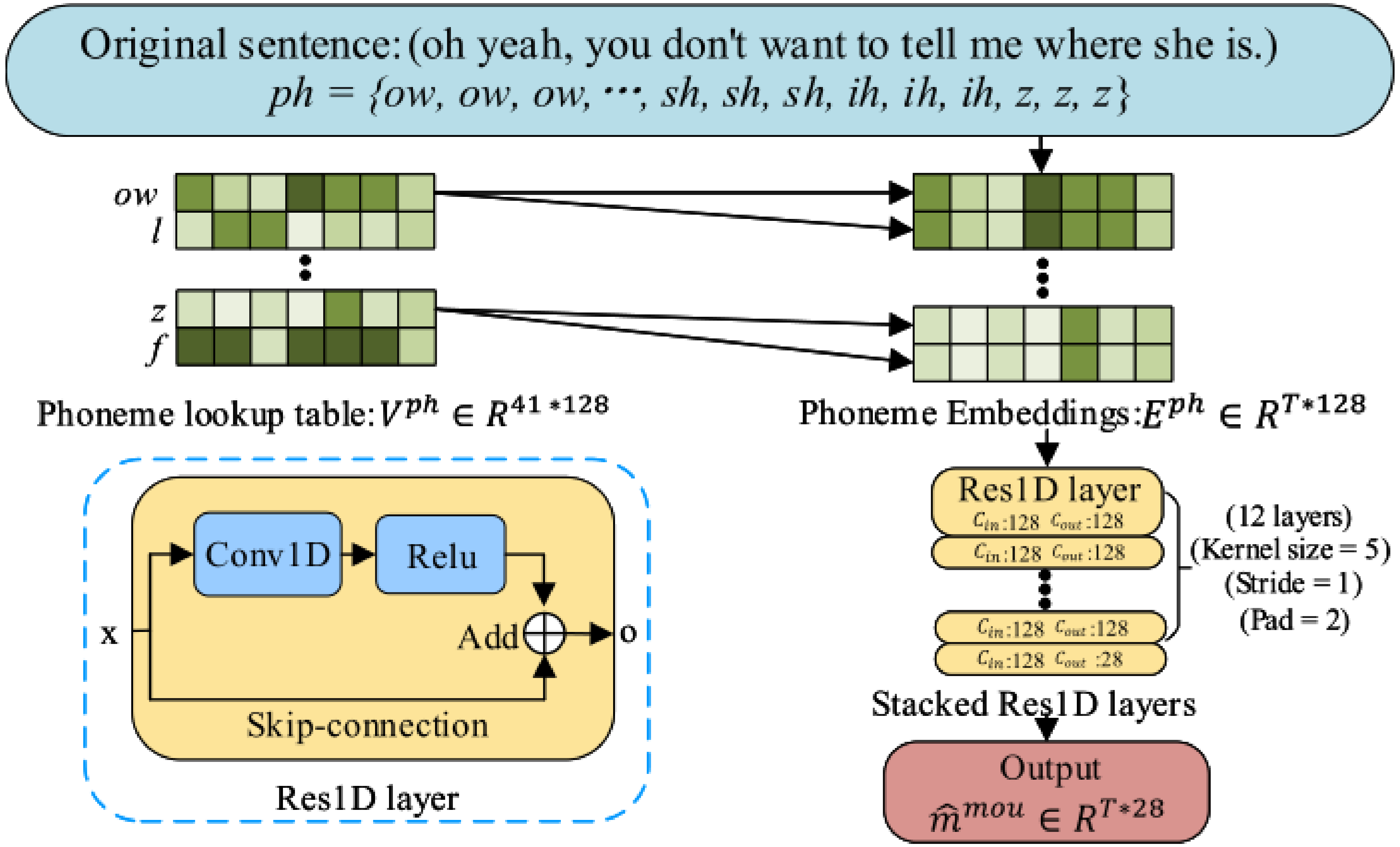


Recurrent model along with frame t



Parallel computing with frames 1, …, T

Neural Network: Lower Facial Expression


Original sentence: (oh yeah, you don't want to tell me where she is.)
$ph = \{ow, ow, ow, \cdots, sh, sh, sh, ih, ih, ih, z, z, z\}$
ow
l
z
f
Phoneme lookup table: $V^{ph} \in R^{41*128}$
Phoneme Embeddings: $E^{ph} \in R^{T*128}$
Res1D layer
C_{in}:128 C_{out}:128
C_{in}:128 C_{out}:128
C_{in}:128 C_{out}:128
C_{in}:128 C_{out}:28
(12 layers)
(Kernel size = 5)
(Stride = 1)
(Pad = 2)
Stacked Res1D layers
Output
$\hat{m}^{mou} \in R^{T*28}$
Conv1D
Relu
x
Add
o
Skip-connection
Res1D layer

Memory: 17M

Training time: 8 hours on GPU 1080ti

Test time: 25-50ms for a 15-second utterance

Output: T X 28 animation sequence

Performance: lower facial expression

Synthetic
Mocap

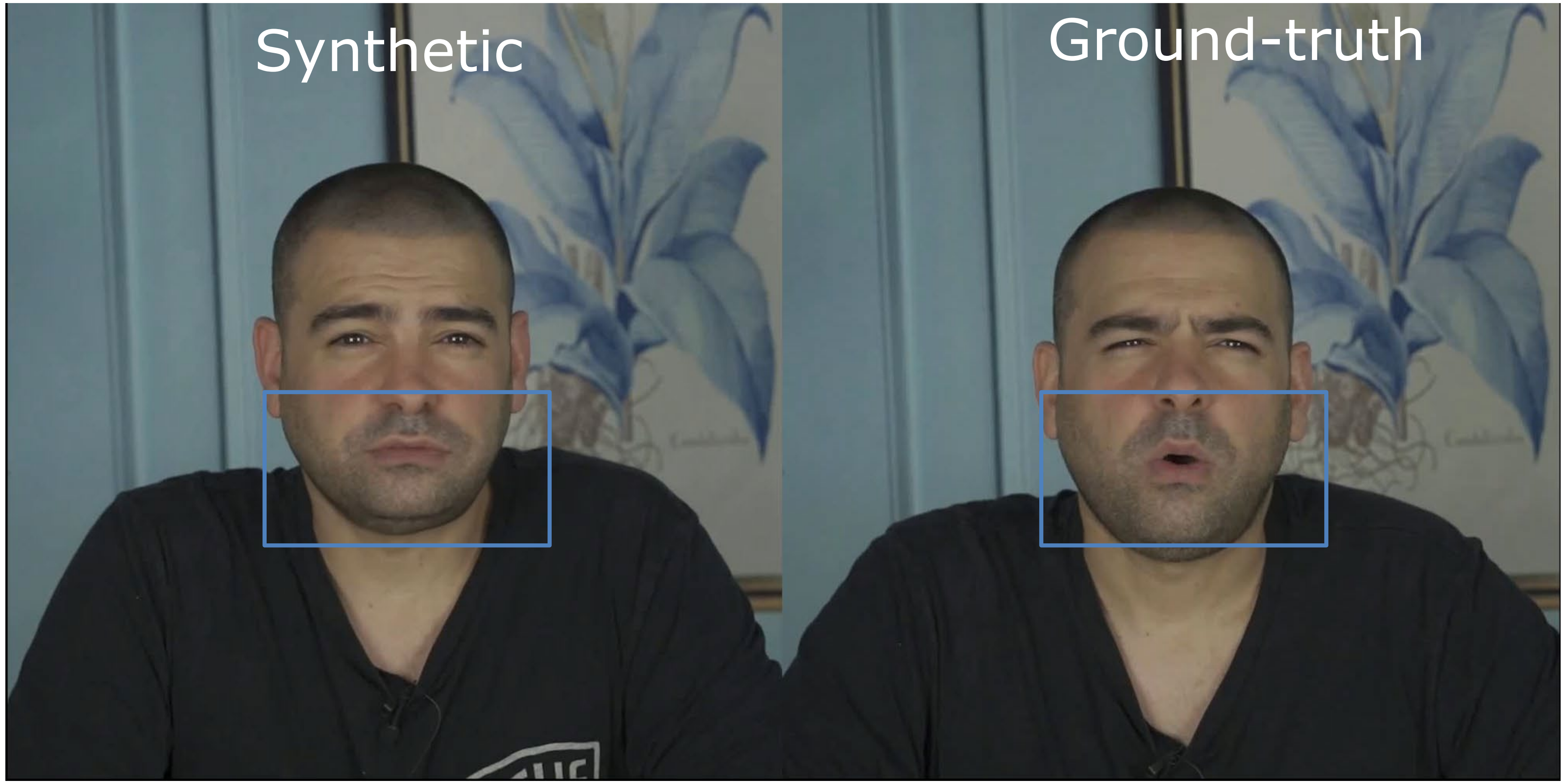
Synthetic
Mocap

Synthetic
Mocap

Synthetic
Mocap

Performance: lower facial expression

Generalized to photo-realistic mouth movement generation



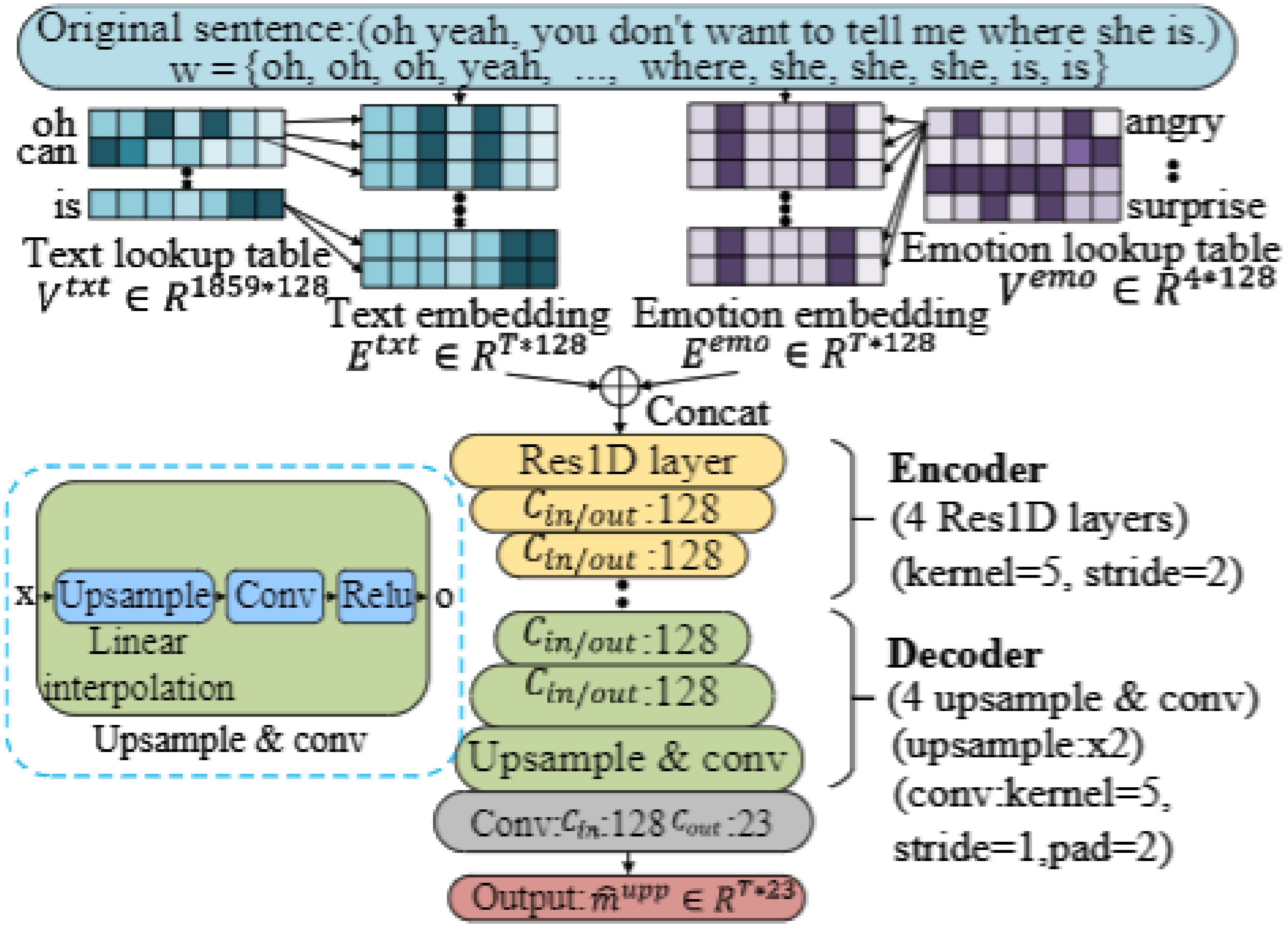
Neural Network: Upper Facial Expression

- Eyelids
- Eyebrow

Neural Network: Upper Facial Expression

- Trick1: using time-aligned words, instead of speech features (e.g. MFCC)
- Trick2: using CNN-based model, instead of LSTM-based ones

Neural Network: Upper Facial Expression



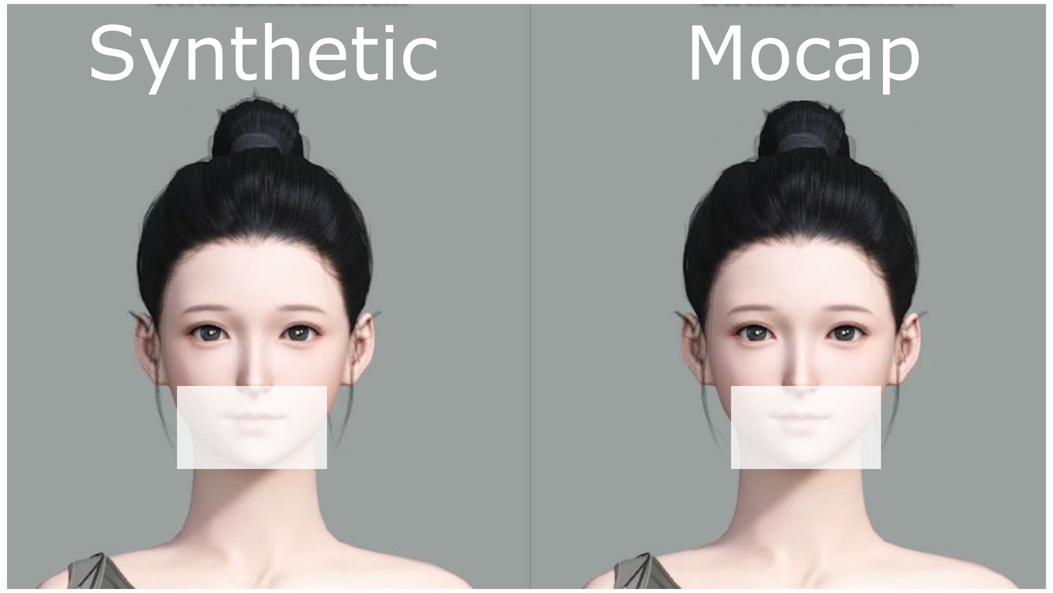
Output: T X 23 animation sequence

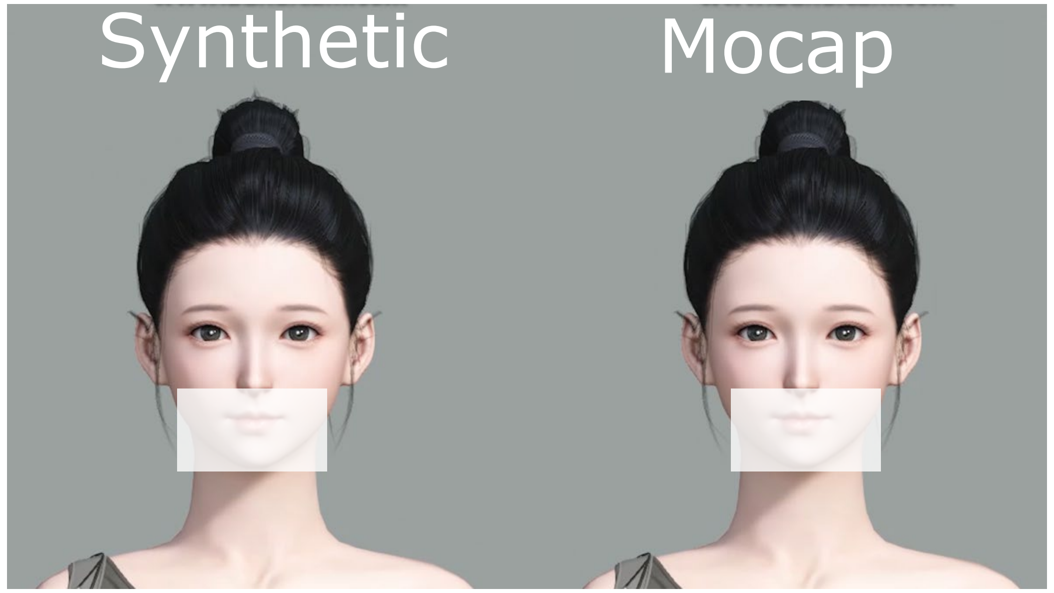
Memory: 13M

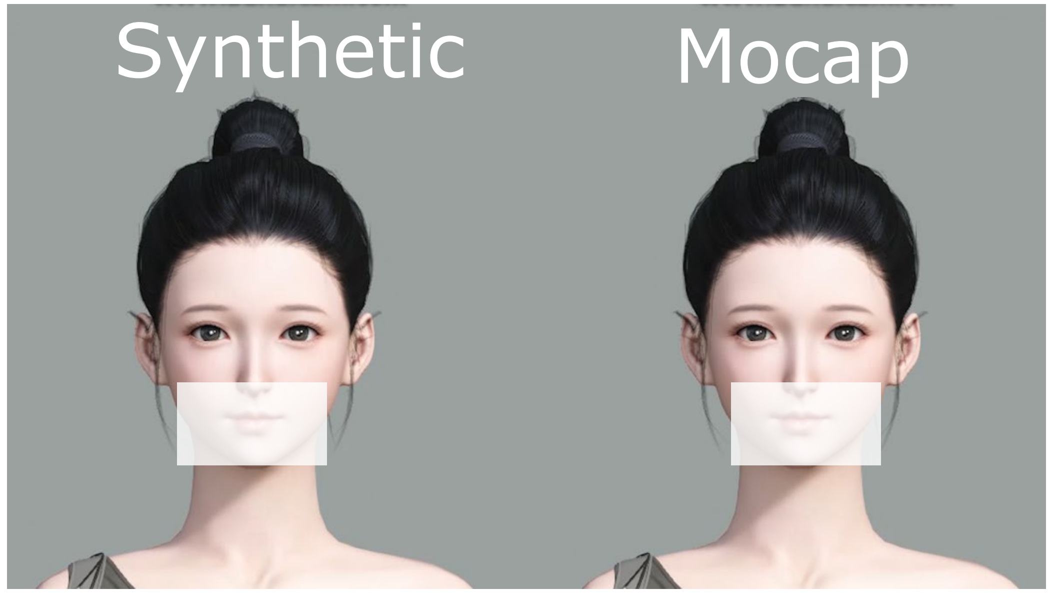
Training time: 3 hours on GPU 1080ti

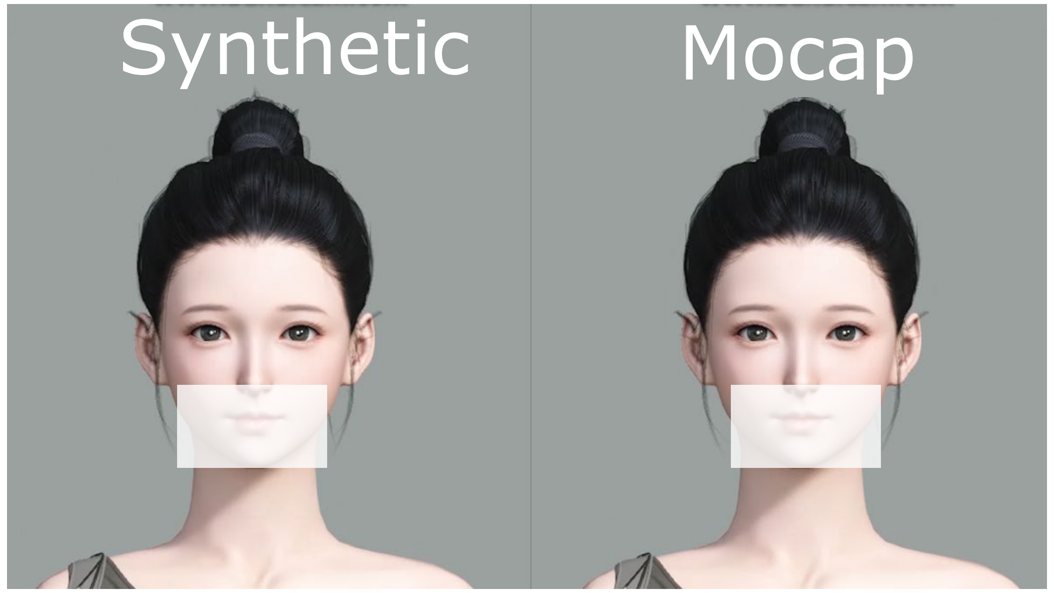
Test time: 25-50ms for a 15-second utterance

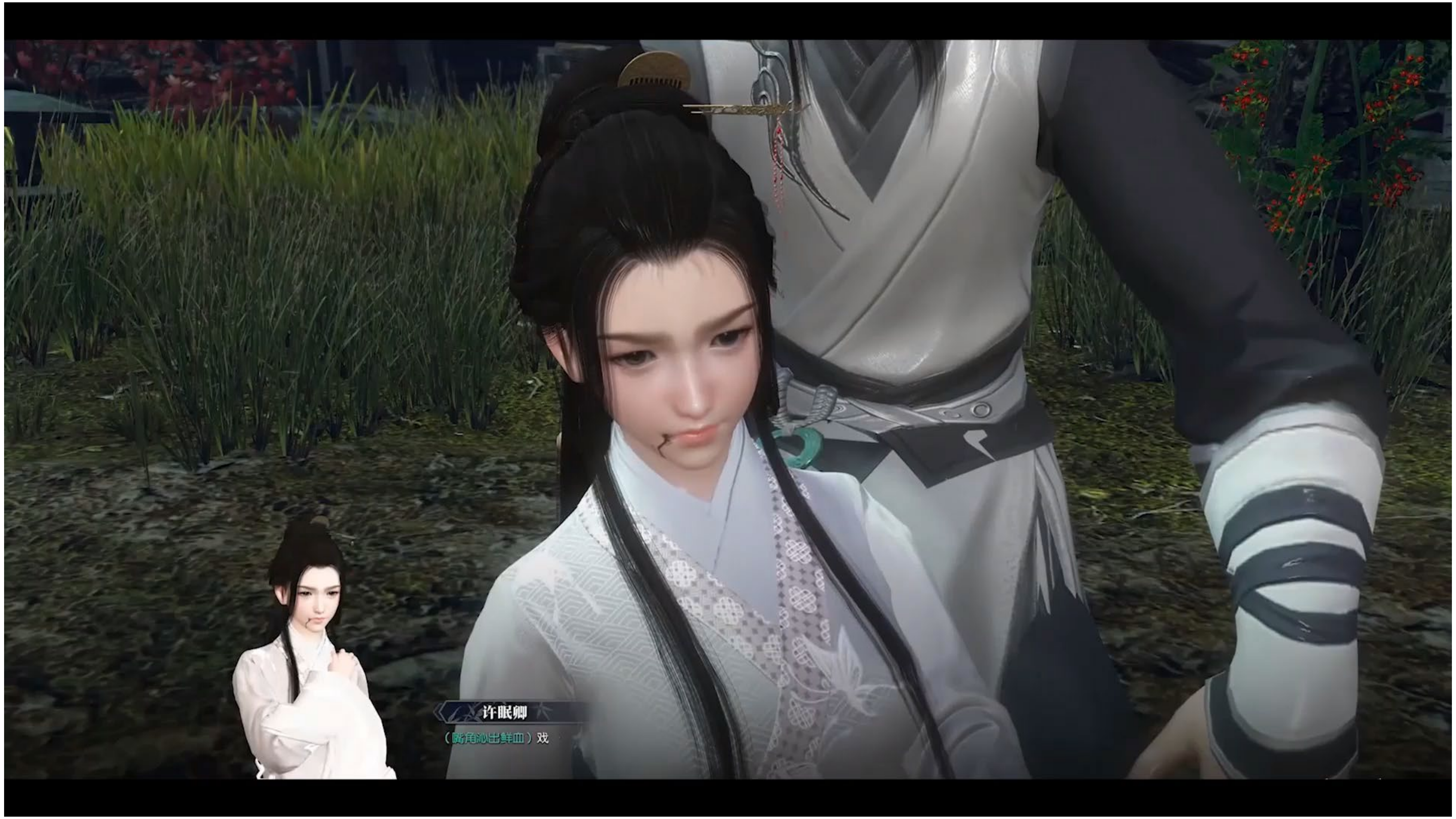
Performance: upper facial expression











许眠卿

（嘴角沁出鲜血）戏

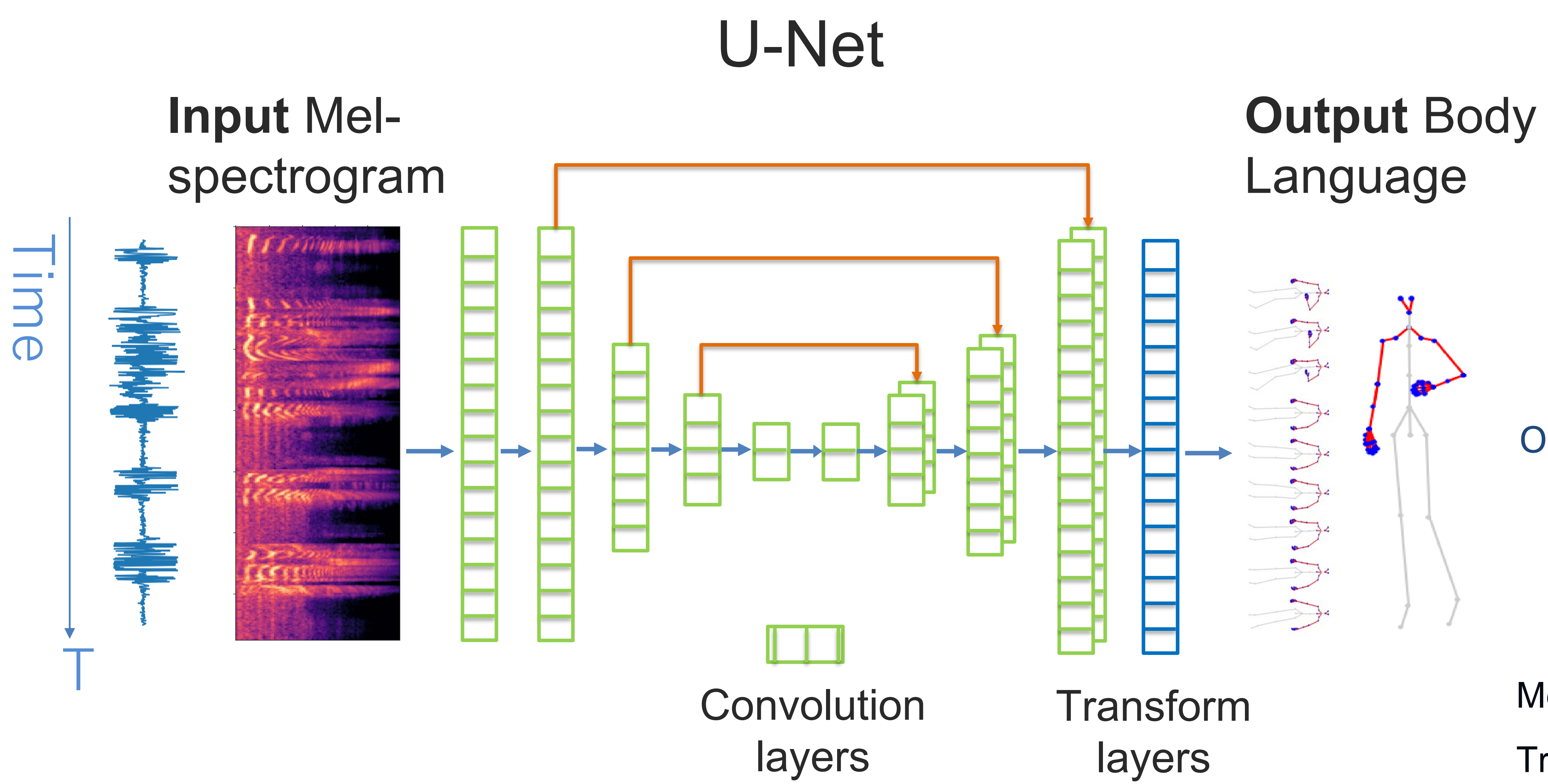
夏颜

......

Neural Network: body language

- Head
- Hand
- Torso
- Legs

Neural Network: body language



- ✓ ***Achieving large receptive field with little computational cost;***
- ✓ ***Better capturing multi-scale features using hierarchical network architecture***

Output: T X 76 animation sequence

Memory: 6.8M

Training time: 3 hours on GPU 1080ti

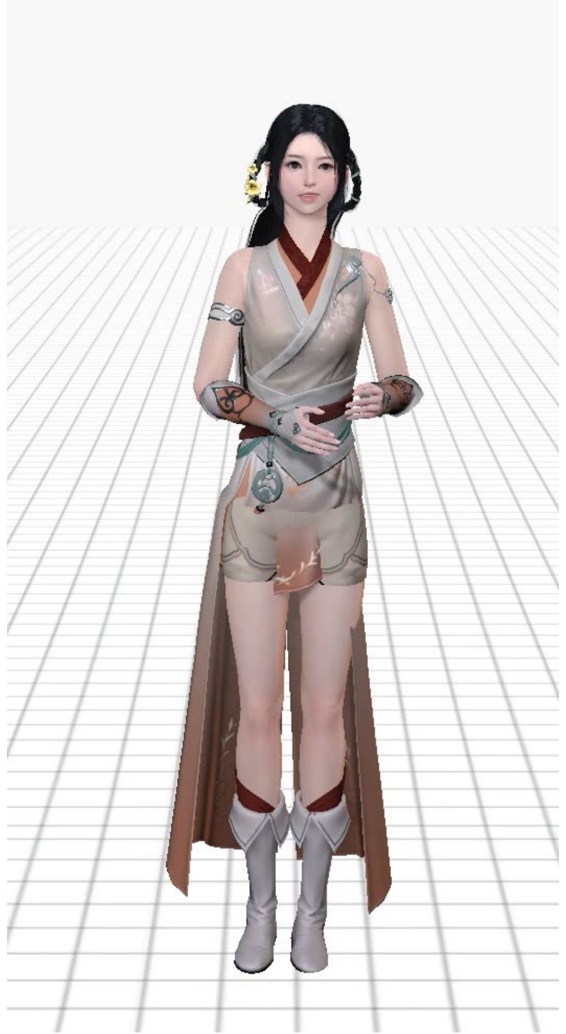
Test time: 300ms for a 15-second utterance

Performance: body language

Neutral

Silly child, the world is too grand, and too complicated.

Happy



How do you know that I love him? Thank you.

Doubt

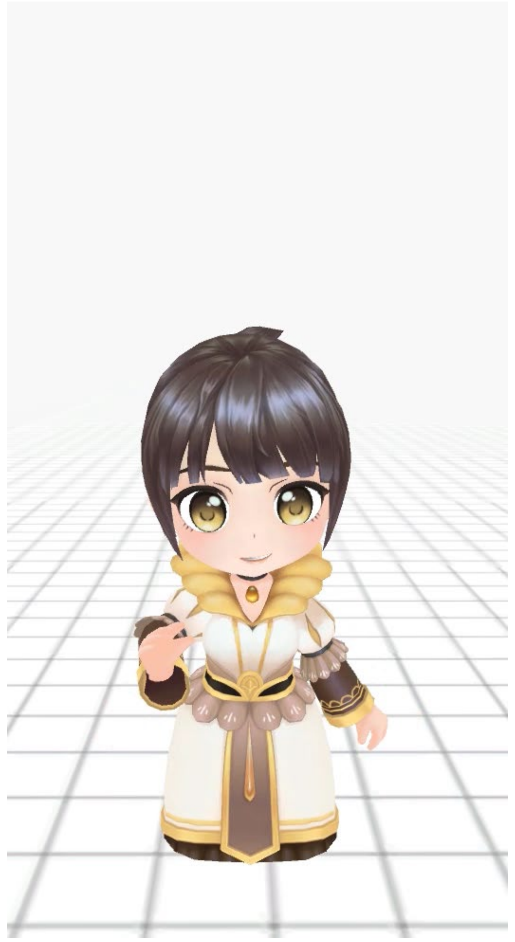
This song voice hears from Miss. Qi. But where is she?

Performance: body language

NPCs express sadness with the same speech.

Ancientry, Introvert

Cartoon, Lively

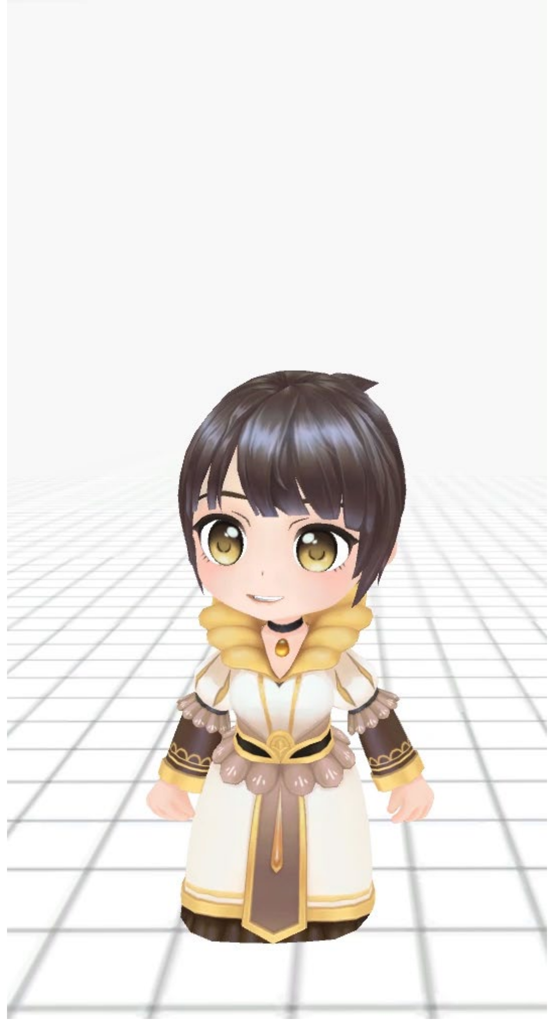


Cartoon, Introvert

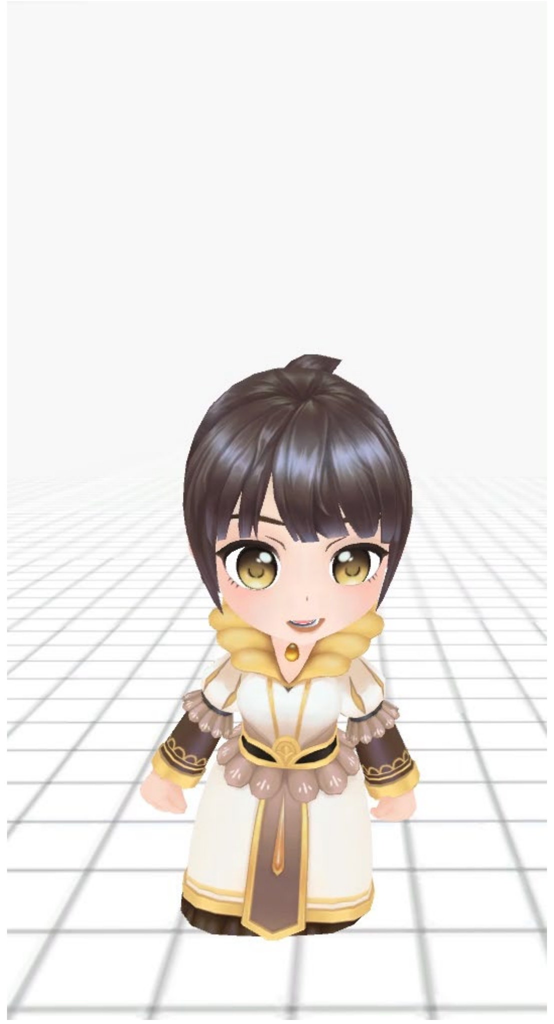
别说了，说的我心里好难受。
Stop it, I feel so upset.

Performance: body language

Neutral



Happy



Angry

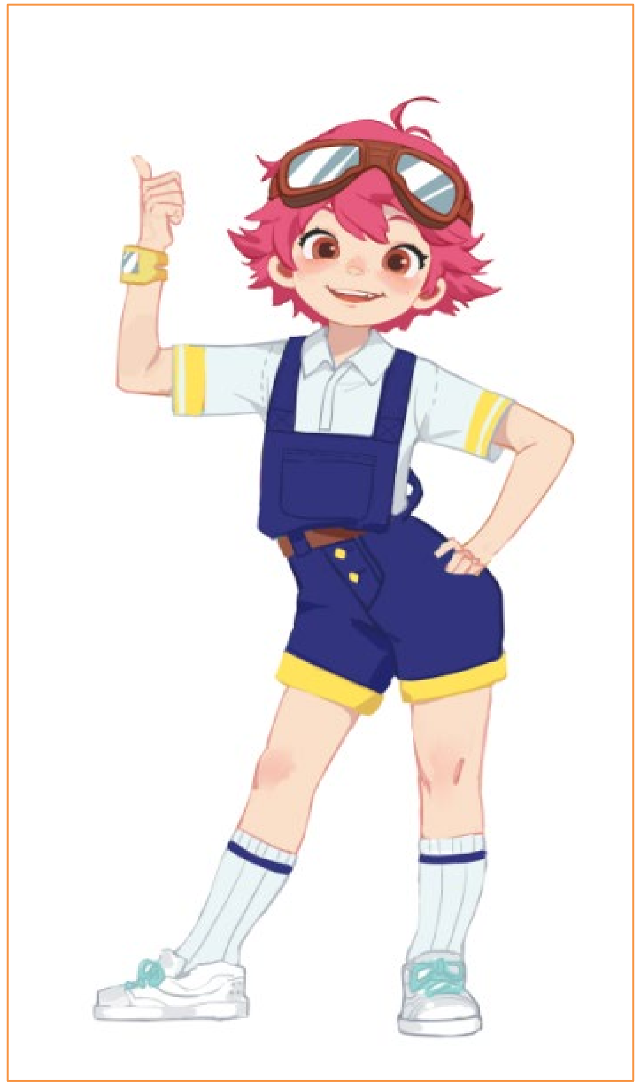
A coding game with more than 12 millions users

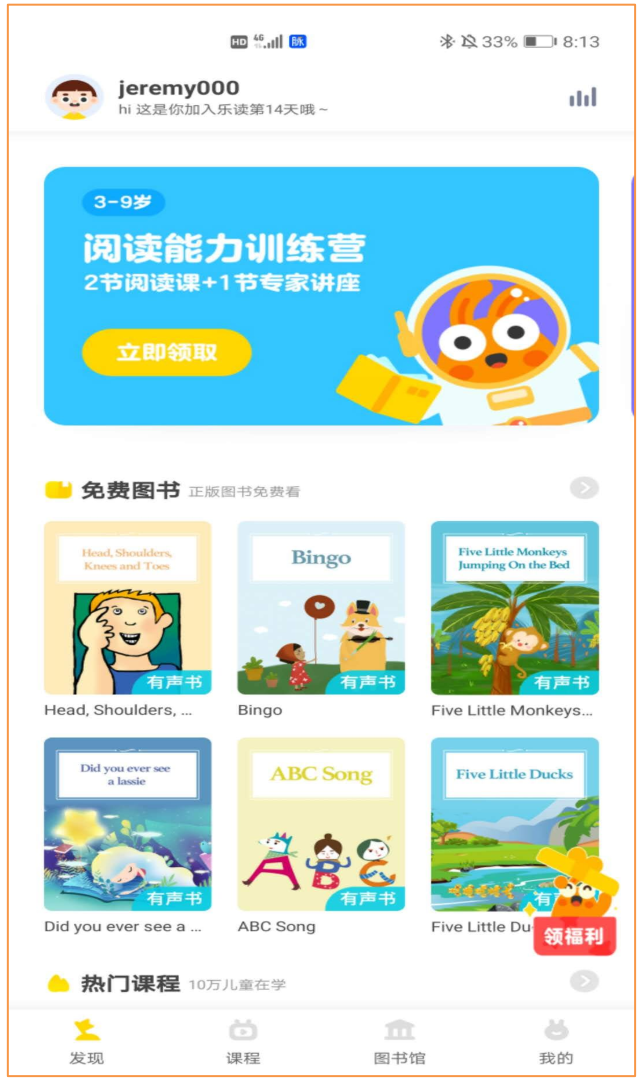
CODE COMBAT

Euler as AI teacher in the game

“In fact, when we are studying, the more we think and ask, the more fun knowledge we will find.”

Performance: body language





8:13 33%

jeremy000

hi 这是你加入乐读第14天哦～

3-9岁

阅读能力训练营

2节阅读课+1节专家讲座

立即领取

免费图书 正版图书免费看

Head, Shoulders, Knees and Toes — 有声书 — Head, Shoulders, ...

Bingo — 有声书 — Bingo

Five Little Monkeys Jumping On the Bed — 有声书 — Five Little Monkeys...

Did you ever see a lassie — 有声书 — Did you ever see a ...

ABC Song — 有声书 — ABC Song

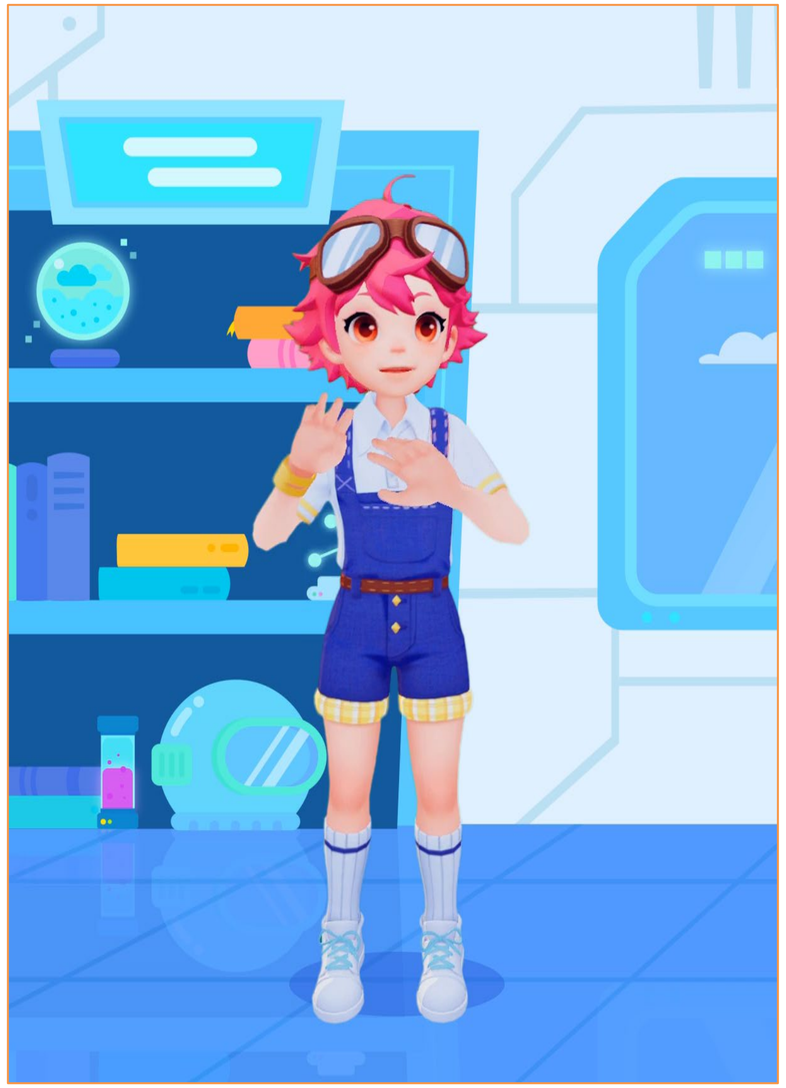
Five Little Ducks — 有 — Five Little Du

领福利

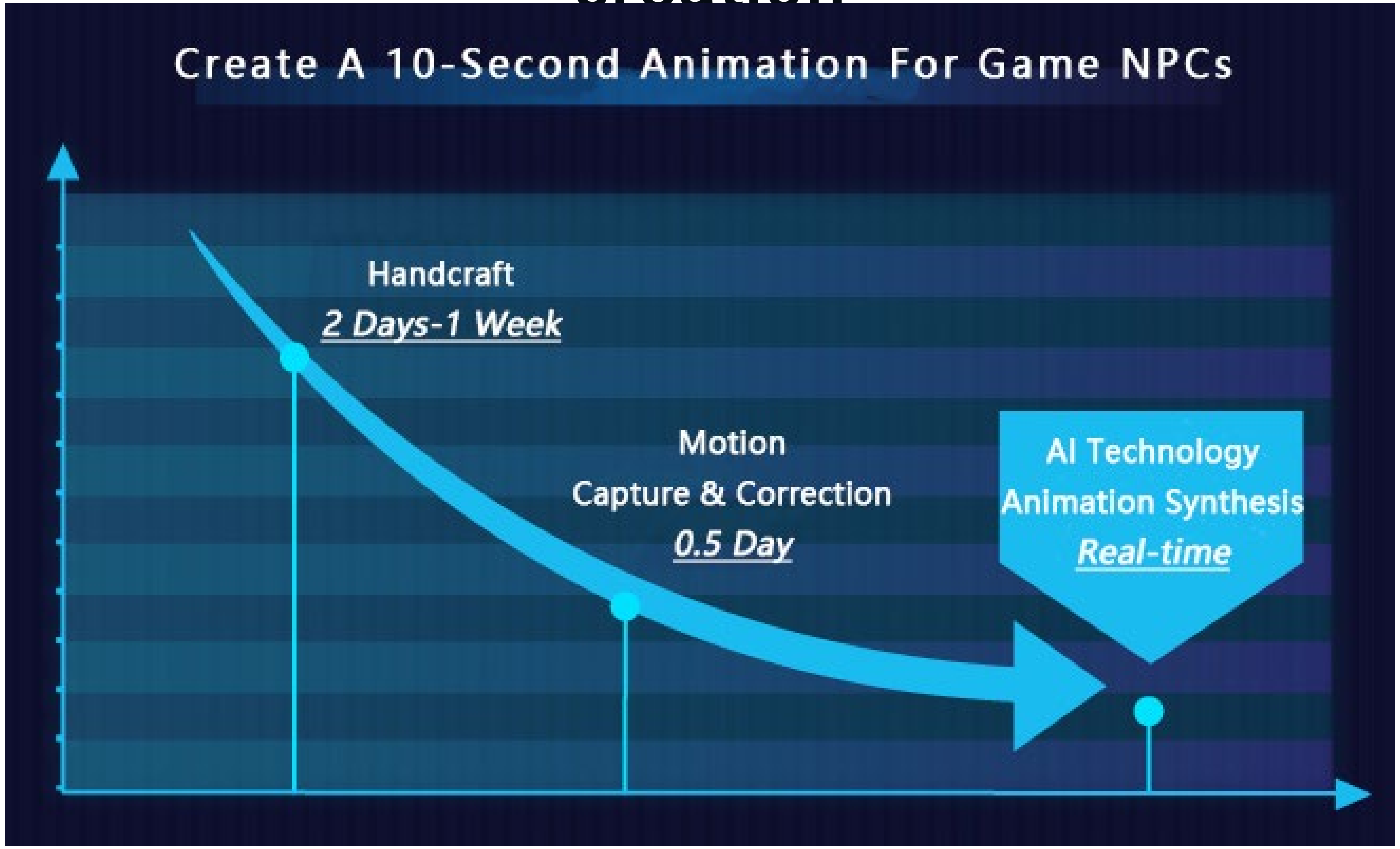
热门课程 10万儿童在学

有道乐读

Happy-reading App



A new era of animation creation



Takeaways

- While speaking, it exists correlations between human behaviors and speech.
- Human behaviors are temporally synchronized with the prosodic and syntactic structure of the speech
- Deep learning techniques can build correlations between human behaviors and simultaneous speech.
- Neural Networks are powerful for the automatic generation of full-body animations according to speech signals.
- With neural networks, the face/body animation generation requires the time-consuming in less than 60ms/600ms.

Limitations

- There is still a long way to go in making the generated animations more expressive.
- There will always be a need to retarget the output animations to NPCs with different skeletons.
- The frameworks require high-quality motion caption data.
- The styles of generated animation depend on the recorded dataset.

Future work

- Effectively taking into account both text and audio information for more believable behavioral expressions.
- Jointly taking into account inherent correlations among three-modality animations.
- Making more semantic facial expressions and body behaviors.