A. Implementation Details

A.1. Video Preprocessing

For video data preprocessing, we employ the facial cropping method introduced in [3]. To further enhance computational efficiency, we standardize the input frame resolution to 96×96 pixels, balancing detail retention with processing speed.

A.2. Audio Preprocessing

In alignment with [3], we generate Mel-spectrograms from audio samples recorded at 16kHz. This approach involves setting a window size of 800 samples and a hop size of 200, ensuring temporal resolution appropriate for capturing the nuances of lip movements.

A.3. Visual Encoder

The Visual Encoder plays a critical role in modeling spatial features from the talking face video. It consists of a deep convolutional architecture with 18 2D convolutional layers, each accompanied by Batch Normalization and ReLU activation functions. In addition, a subset of these modules incorporates residual blocks and skip connections, enhancing its ability to model complex visual features.

A.4. Global Emotion Text Encoder

The Global Emotion Text Encoder leverages a pretrained Emoberta [2] model to encode the overarching emotional tones within the caption text. We import *SentenceTransformer* from the *sentence_transformers* library in Python and specifically use the *tae898/emoberta-base* model.

A.5. Linguistic Text Encoder

We adopt a pretrained GPT-Neo [1] model as our Linguistic Text Encoder. The encoder is set up using the *transformers* library in Python. We import *AutoTokenizer* and *GPTNeoModel* from *transformers* library. The tokenizer is initialized with *AutoTokenizer.from_pretrained("EleutherAI/gpt-neo-2.7B")*, and the model is initialized with *GPTNeoModel.from_pretrained("EleutherAI/gpt-neo-2.7B")*, respectively.

A.6. Visual Decoder

The Visual Decoder is composed of six sets of 2D transpose convolution blocks, with each block consisting of a 2D transpose convolution layer and two layers of 2D residual convolution blocks. This architectural design enables the transformation of text descriptions into a coherent and expressive sequence of video frames.

A.7. Loss Weight Tuning

λ_1, λ_2, and λ_3 are the weight of three losses, fine-tuned for optimal performance. During the initial 300 epochs, we assign 0.7 to λ_1, 0.09 to λ_2, and 0.21 to λ_3. Afterward, we adjust these weights, setting λ_1 to 0.9, λ_2 to 0.03, and λ_3 to 0.07.

B. Loss Structural Insights

We explore the structural aspects of the Face Synthesizer and the Discriminator, highlighting their respective contributions and characteristics.

The Face Synthesizer in FT2TF is inspired by the frozen-weight, pretrained Syncnet model [3]. It serves as a critical component in the lip synchronization process by aligning the synthesized talking face frames with the corresponding audio data. While not explicitly detailed through mathematical expressions, the Face Synthesizer integrates audio information to enhance the naturalness of lip movements in the generated talking face frames.

The Discriminator is composed of 14 layers of fully convolutional modules, without the use of Batch Normalization layers or skip connections. It operates as a binary classifier, providing binary predictions (O_{pred}) based on whether the input is generated or Ground Truth. These predictions are guided by the binary labels (Y_{disc}). This binary classification loss, as detailed in L_{disc}, helps to ensure the quality and authenticity of the synthesized frames.

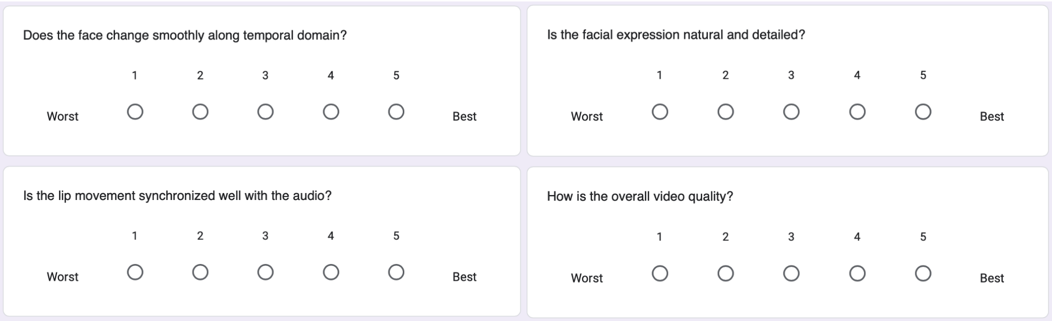
In summary, FT2TF's loss functions, coupled with the structural characteristics of the Face Synthesizer and the Discriminator, facilitate a comprehensive training process. These components work in harmony to improve the pixel-level fidelity, lip synchronization, and overall realism of the generated talking face videos, making FT2TF an effective solution for natural and expressive talking face synthesis.

C. User Study

To evaluate the quality of FT2TF's generated talking faces in comparison to existing methods, we conduct a comprehensive user study assessing participant impressions across multiple dimensions.

Questionnaires. The questionnaire used in the user study is shown in Figure S1. Four questions are designed for each method, covering key aspects: temporal transition smoothness, lip synchronization accuracy, facial detail naturalness, and overall video quality.

User Responses. Figure S2 provides a summary of user responses. The results indicate that participants responded significantly more positively to FT2TF compared to other methods.



Does the face change smoothly along temporal domain?

	1	2	3	4	5	
Worst	○	○	○	○	○	Best

Is the facial expression natural and detailed?

	1	2	3	4	5	
Worst	○	○	○	○	○	Best

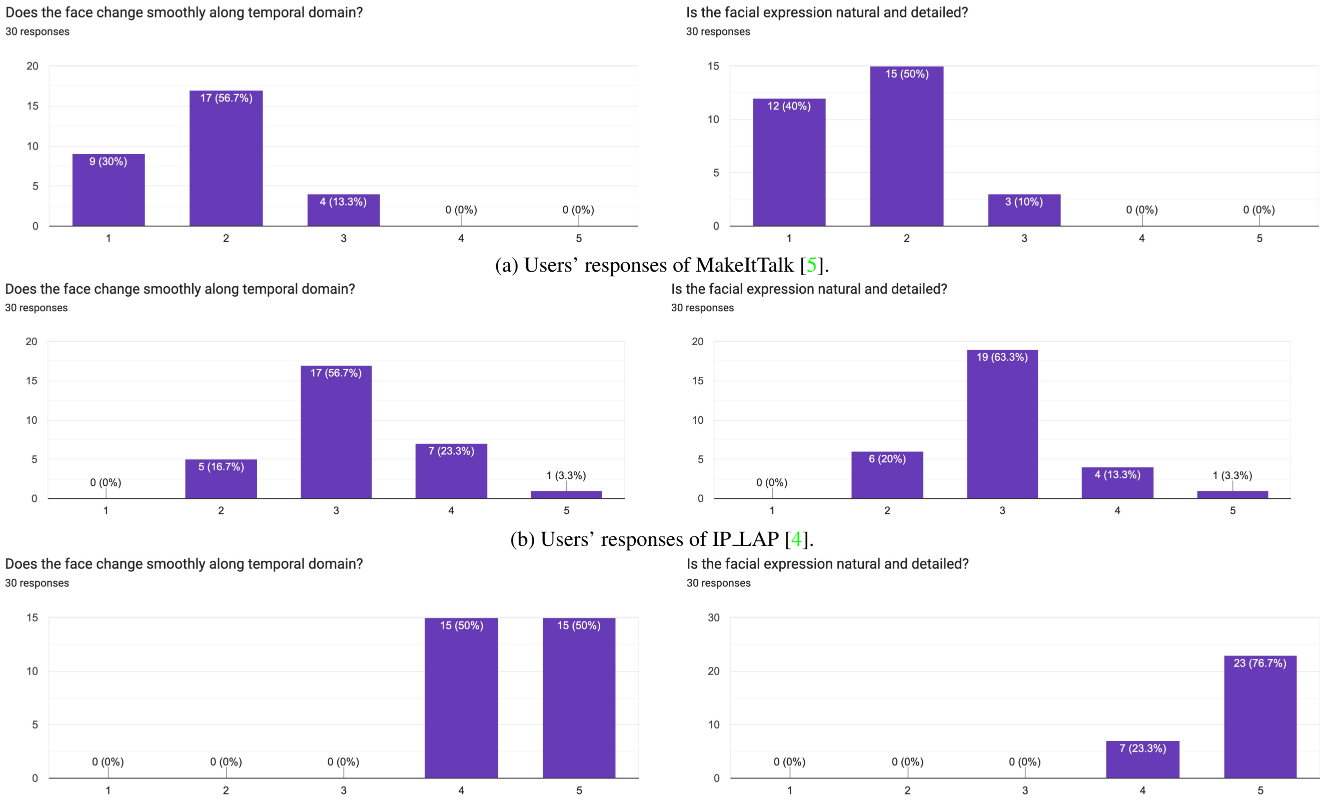
Is the lip movement synchronized well with the audio?

	1	2	3	4	5	
Worst	○	○	○	○	○	Best

How is the overall video quality?

	1	2	3	4	5	
Worst	○	○	○	○	○	Best

Figure S1. **User study questions.** We design four questions for each method in the user study to evaluate the quality of generated talking faces across multiple dimensions.



(a) Users' responses of MakeItTalk [5].

(b) Users' responses of IP_LAP [4].

(c) Users' responses of **FT2TF (ours)**.

Figure S2. **Summary of user responses across methods.** The results indicate that user evaluations are notably more favorable for FT2TF compared to other methods.

References

[1] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 1

[2] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021. 1

[3] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *International Conference on Multimedia*, 2020. 1

[4] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Conference on Computer Vision and Pattern Recognition*, 2023. 2

[5] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevar-

ria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *Transactions On Graphics*, 2020. 2