Research Documentation

1 Human Talking Head Synthesis

Human Talking Head synthesis includes two significant tasks:

1. Motion transfer from a driving video to the source image.
2. Lip-synchronization of the motion transferred video with the speech generated on the other hand.

1.1 Motion Transfer Models

To identify the most suitable model for motion transfer, I conducted extensive research and experiments with several models. Here are the key models explored:

- **Thin-Plate Spline Motion Model** and **First-Order Motion Model**: These were among the earliest models tested. They offered superfast inference times but lacked accuracy and high-resolution output. The resulting videos were often jittery and inconsistent.

- Thin-Plate Spline Motion Model
- First-Order Motion Model

- **Latent Image Animator**: This model provided consistent and fast inference but did not produce high-resolution or sharp output. Although considered initially, we later decided to prioritize sharper output over inference time, leading us to look for other models.

- Latent Image Animator

- **AdaSR Talking Head**: This model produced sharper outputs compared to Latent Image Animator. However, it still did not achieve megapixel-range resolution.

- AdaSR Talking Head

- **LivePortrait**: A newer model with impressive high-resolution output. However, its motion transfer consistency was insufficient for our specific talking head purposes.

- LivePortrait

Given the trade-offs, we finalized the **AdaSR model** for motion transfer, achieving a resolution of 512x512, which is compatible with MuseTalk lip synchronization output.

1.2 Lip-Synchronization Models

For the lip synchronization task, I evaluated several models, focusing on achieving low inference times while maintaining high quality:

- **Wave2Lip**: An early model known for its good lip-sync performance and fast inference. However, it produced poor sharpness and low-resolution output. Attempts to enhance it using Real-ESRGAN resulted in increased inference time and loss of facial details.

- Wave2Lip
- Real-ESRGAN

- **SadTalker**: An end-to-end model that performed well in Mandarin but not in English. It also required an enhancer like Real-ESRGAN, which increased inference time.

- SadTalker

- **GeneFacePlusPlus**: This model required training on person-specific videos to store appearance volumetric features. While promising, it necessitated a large set of curated videos for effective performance.

- GeneFacePlusPlus

- **AniPortrait**: Provided excellent results but had an extremely high inference time, making it unsuitable for our needs.

- AniPortrait

- **MuseTalk**: A VAE-based lip-synchronization model that delivered consistent results across different languages. It allowed storing appearance-volumetric features (avatars) and generated talking heads at approximately 30fps on an RTX4080 GPU and 25fps on V100 and A100 GPUs. This model is well-suited for live streaming, requiring only 25 frames per second to achieve a consistent 1-second video.

- MuseTalk

2 Animated Talking Head Synthesis

To achieve animated talking head synthesis, I explored two primary approaches:

2.1 Custom Pipeline Creation

This approach involved using various models to create a pipeline, although the inference time was uncertain:

- **LiveWhisper** based on OpenAI's Whisper for transcription:

- LiveWhisper

- **Groq** for large language model output:

- Groq

- **AllTalkTTS** or **StyleTTS2** for text-to-speech synthesis:

- AllTalkTTS

- StyleTTS2

- **Talking-Head-Anime-3** for animation:

- Talking-Head-Anime-3

- **EasyVtuber** for further animation and control:

- EasyVtuber

- **SoftVC VITS** (optional) for making the character sing:

- SoftVC VITS

Due to time constraints and the promising alternative approach, I did not create a pipeline using these models. However, this remains a potential future endeavor.

2.2 SillyTavern-Extras Pipeline

The SillyTavern-Extras extension, utilizing Tha3 models, promised 30fps generation on advanced GPUs, suitable for live streaming:

- This extension animates the character's expressions based on API calls but did not initially account for audio-based mouth animations. - I modified the pipeline to map mouth animations with phonemes extracted from the audio file using a multilingual wave2vec2-Espeak model. This modification resulted in the animator stopping mouth animations during audio pauses and synchronizing lip movements with the speaker's pace.

Links to the models used:

- SillyTavern-Extras (Modified work is available in the submitted git repository)

- Wave2Vec2 phoneme extraction model

3 Text-to-Speech Synthesis

I experimented with various TTS models to find the most suitable ones for both human and animated talking heads:

3.1 TTS Models for Human Talking Heads

- **TextrolSpeech**: This model provided good emotional separation but had poor quality output.

- TextrolSpeech

- **OpenVoiceV1** and **OpenVoiceV2**: These models were good but not the best I found later.

- OpenVoice

- **Coqui TTS**: Excellent in terms of quality and inference but had dependency issues with our pipeline.

- Coqui TTS

- **StyleTTS2**: The best model for quality and compatibility with our pipeline. Currently supports only English but can be trained on other languages like Japanese.

- StyleTTS2

3.2 3.2 TTS Models for Animated Talking Heads

- **AllTalkTTS**: Based on Coqui TTS with no dependency issues. Supports multiple languages, including Japanese and Hindi.

- AllTalkTTS

3.3 Discussion on Hindi TTS

I believe training StyleTTS2 on a "Hinglish" dataset is more advantageous than using existing models for Hindi TTS. Current models like AllTalkTTS and Coqui TTS are based on the Devanagari script, adding complexity with text aligners supporting Devanagari. Most Hindi-speaking users, fortunately, or unfortunately, prefer "Hinglish" (Hindi using the English alphabet), making it easier to train and generate better speech outputs without needing a new text aligner. Also, it caters to the best of Hindi-speaking users regarding their preferences.

4 VASA-1 Implementation

VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real-Time is a research initiative from Microsoft for generating real-time talking heads. It comprises two main parts:

4.1 Expressive and Disentangled Face Latent Space Construction

Based on the "MegaPortraits: One-shot Megapixel Neural Head Avatars" paper by Samsung: - I implemented some architectures from scratch and cherry-picked others from the "MetaPortrait" implementation.

- MetaPortrait

- Video processing before training required background removal, which I adapted from other implementations like "Deep3DFacereconstruction".

- Deep3DFacereconstruction

- Additional architectures from papers like "VOODOO 3D" and "Arcface from Insightface" were liberally implemented.

- VOODO3D

- Insightface

4.2 Holistic Facial Dynamics Generation with Diffusion Transformer

The VASA-1 paper provided vague details on the diffusion transformer. To implement it: - I referred to papers like "Scalable Diffusion Models with Transformers (DiT)" and "DiffPoseTalk: Speech-Driven Stylistic 3D Facial Animation and Head Pose Generation via Diffusion Models".

- DiT
- DiffPoseTalk

Although VASA-1 was implemented, training will be deferred due to the time-intensive and liberal implementation resulting from the paper's vague descriptions. Future improvements will involve reviewing soon-to-be-released papers like "EMOPortraits" and "EMO: Emote Portrait Alive".

- EMOPortraits
- EMO: Emote Portrait Alive

5 Conclusion

This document details the comprehensive research and experiments conducted to achieve human talking head synthesis, animated talking head synthesis, text-to-speech synthesis, and VASA-1 implementation. The models and links provided offer a robust foundation for future improvements and implementations in this domain.