# AIC Final Project
# Image-Driven Facial Animation with PaddleGAN: A Web-Based System and Comparative Study on Face Angles and Styles

Team：B6
111550040 曾紹幃 111550143 林彥佑 111550149 林悅揚

Contributions of the individual group members
曾紹幃(30%)：Experiments, Analyze, Report, PPT
林彥佑(30%)：Model Testing, Analyze, Report, PPT
林悅揚(40%)：Model Testing, Experiments, Analyze, Report, PPT

# 1. Motivation

With the rising popularity of AI-based face animation and face-swapping technologies, we were motivated to create a facial animation tool that is accessible to everyone, aiming to lower the technical barrier for general users.

We implemented our system using PaddleGAN's First-Order Motion Model, but also observed that its performance can vary significantly across different types of input, such as side-view images or stylized illustrations.

Rather than merely building a working prototype, we conducted qualitative evaluations and analyses to better understand both the strengths and limitations of this model under diverse scenarios.

# 2. Techniques used

## A . Core Animation Model: PaddleGAN's First-Order Motion Model (FOMM)

- **Keypoint Detection:**

  We utilize the model's built-in keypoint extractor to detect critical facial and body landmarks in both the source image and the driving video. These keypoints serve as anchors for modeling motion and deformation across frames.

- **Motion Estimation & Spatial Warping:**

  The model estimates a dense motion field based on the relative movement of keypoints between frames. It then applies spatial warping to the source image to create the illusion of motion. This process enables complex transformations such as head rotation, blinking, and subtle 3D-aware deformations—all from a single input image.

- **Occlusion Handling:**

  A dedicated occlusion mask is learned during inference to predict which parts of the face may become hidden or reappear due to movement. This helps maintain visual

coherence and prevents unnatural artifacts from appearing in the generated frames.

## B . Frontend Implementation:

- **React + Vite:**

  We built a modern and responsive frontend using React and Vite, enabling fast development and seamless hot-reloading. The component-based structure improves code reusability.

- **TypeScript:**

  Type safety and better tooling support are provided through TypeScript, making the codebase more robust and maintainable in the long term.

- **Axios:**

  For communicating with the backend API, Axios is used to send uploaded image/video data and retrieve the generated animation results.

- **Tailwind CSS:**

  A utility-first CSS framework that allowed us to quickly build a clean, responsive, and aesthetically pleasing user interface without writing verbose custom styles.

## C . Backend Implementation:

- **FastAPI:**

  We chose FastAPI for its speed, simplicity, and excellent support for asynchronous requests. It handles file uploads, initiates inference, and serves the final result.

- **FFmpeg:**

  Before inference, FFmpeg is used to decode the driving video and extract individual frames at a fixed frame rate. This is essential for feeding consistent temporal data into the model.

- **Python Inference Pipeline:**

The animation process is executed through a Python backend that loads the pre-trained PaddleGAN model and runs inference on the source image and driving frames, generating an output video.

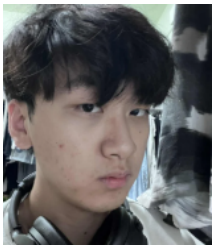
## 3. Experiments Done

To evaluate the performance and limitations of our image-driven facial animation system, we designed a set of qualitative experiments focused on varying the input types. Since the First-Order Motion Model (FOMM) used in PaddleGAN lacks built-in quantitative evaluation metrics (e.g., PSNR, FID, LPIPS), we adopted a **visual comparison approach** based on **human perceptual judgment**.

Our experimental design focused on how **different types of source images and driving videos** influence the quality of the generated animation. Specifically, we selected:

**3 types of source images**:

1. **Frontal face photo** — a standard, well-lit portrait
2. **Side face photo** — profile view to test spatial generalization
3. **Stylized/animated face** — to simulate illustration or cartoon use cases

| Animation | Frontal | Side |
| --- | --- | --- |
|  |  |  |

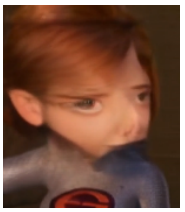
**3 types of driving videos** *(see demo video for examples)*:

1. **Frontal face video with head motion** — typical webcam-style video
2. **Side face video with head motion** — introduces rotational complexity
3. **Animated/cartoon face video** — tests performance on stylized motion

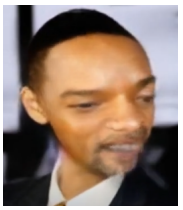
| Animation | Frontal | Side |
|---|---|---|
|  |  |  |

# 4. Results and analysis

### 1 . Animated Face Image with

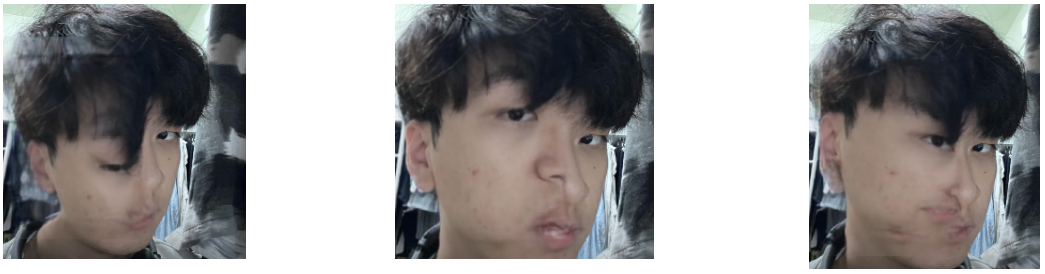
| Animation Video | Frontal Video | Side Video |
|---|---|---|
|  |  |  |

### 2 . Frontal Face Image with

| Animation Video | Frontal Video | Side Video |
|---|---|---|
|  |  |  |

**2 . Side Face Image with**

**Animation Video** **Frontal Video** **Side Video**



| Source Image Type → Driving Video Type ↓ | Animated Video | Frontal Face Video | Side Face Video |
|---|---|---|---|
| **Animated Face Image** | ❌ | ✅ | ❌ |
| **Frontal Face Image** | ❌ | ✅ | ❌ |
| **Side Face Image** | ❌ | ⚠️ | ❌ |

(*Legend: ✅ = Good, ⚠️ = Acceptable but flawed, ❌ = Poor/distorted output*)

Each test case was conducted by uploading the respective source image and driving video through our web-based interface, generating the animation, and recording observations on the **realism**, **stability**, **identity preservation**, and **artifact presence**.

We used the following criteria during visual evaluation:

- **Naturalness of motion**
- **Facial structure preservation**

- **Jitter, tearing, or unnatural warping**
- **Temporal consistency between frames**

By using this systematic setup, we were able to isolate how input angle and style affect the final animation results, laying the foundation for analysis and future improvements.

# 5. Conclusions to Draw from Your Results

Our experimental findings reveal clear patterns in how input types affect the animation quality produced by PaddleGAN's First-Order Motion Model. The model performs best under ideal conditions—specifically when both the source image and the driving video are **frontal, real-world human faces**. In this setting, the generated animation is not only stable but also visually realistic, preserving the identity and facial dynamics effectively.

Surprisingly, **stylized or illustrated source images** yielded relatively good results when paired with a frontal driving video. Despite lacking real facial textures or depth cues, these images were animated in a manner that closely approximated real photographs, demonstrating the model's adaptability to some artistic styles.

However, the model **struggles significantly with side-view inputs**—whether as the source image or driving video. In such cases, we observed clear distortions, unnatural warping, and reduced stability. This suggests a fundamental limitation in the model's ability to handle non-frontal viewpoints, likely due to insufficient 3D spatial reasoning and keypoint accuracy in those scenarios.

Additionally, **animated driving videos** tended to produce poor motion transfer regardless of the source image quality. The movement often lacked natural flow, resulting in jerky or rigid animations. This indicates that the model is more sensitive to the quality and realism of the motion in the driving video than to the source image itself.

In summary, while the First-Order Motion Model is effective in animating frontal, realistic images, its performance degrades sharply when applied to side views or stylized motions. This underlines the need for enhanced 3D-awareness, more robust keypoint detection, and possibly model fine-tuning for specific domains such as cartoon or illustration inputs.

# 6. What You Have Learned from the Project

### 1.Input Alignment Matters:

We learned that input alignment and camera viewpoint (frontal vs. side) play a critical role in determining the animation's realism and stability. Proper alignment results in more coherent and natural-looking motion.

### 2.Frontal Inputs Work Best:

Frontal source images and driving videos consistently produce the most stable and visually convincing results. In contrast, side-view or misaligned inputs often lead to facial distortions, jitter, or motion artifacts.

### 3.Stylized Inputs Have Limitations:

The First-Order Motion Model shows clear performance limitations when handling stylized or cartoon-like images, particularly when paired with animated driving videos. These combinations tend to break temporal coherence and realism.

### 4.Lack of Quantitative Evaluation Standards:

We observed that the absence of standardized metrics for evaluating facial animation quality necessitates human-centric evaluation. This exposes a broader challenge in benchmarking generative motion transfer models.

### 5.Practical Full-Stack Development Experience:

From a systems engineering perspective, we gained hands-on experience building a user-friendly web platform by:

- Developing a modern React + TypeScript frontend using Tailwind CSS for clean, responsive design

- Implementing a robust FastAPI + Python backend for file handling, video processing (with FFmpeg), and AI inference

- Ensuring seamless frontend-backend integration to enable smooth, end-to-end facial animation from upload to preview and download

## 7. Future Work

To improve generalizability and animation quality, we propose the following directions:

- Enhance 3D spatial awareness to better handle side views and depth variation
- Improve occlusion handling and keypoint stability to reduce facial distortion and jitter
- Introduce style-adaptive training to support cartoon or illustration-style inputs more effectively
- Add interactive features like motion segment selection or background replacement, to boost user control and creativity

## 8. Reference

PaddleGAN:
https://github.com/PaddlePaddle/PaddleGAN/tree/develop

First-Order-Model:
https://github.com/AliaksandrSiarohin/first-order-model?tab=readme-ov-file

Material provided by Prof. Liu, Yu-Lun:
https://e3p.nycu.edu.tw/pluginfile.php/274689/mod_folder/content/0/20250407_yulunliu_3D_Vision.mp4