F5-TTS

1) **Giới thiệu**

F5-TTS là mã nguồn chính thức cho nghiên cứu "F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching". Dự án này nhằm phát triển một hệ thống chuyển văn bản thành giọng nói (Text-to-Speech) tự nhiên và chính xác bằng cách sử dụng kỹ thuật Flow Matching.

2) **Công nghệ**

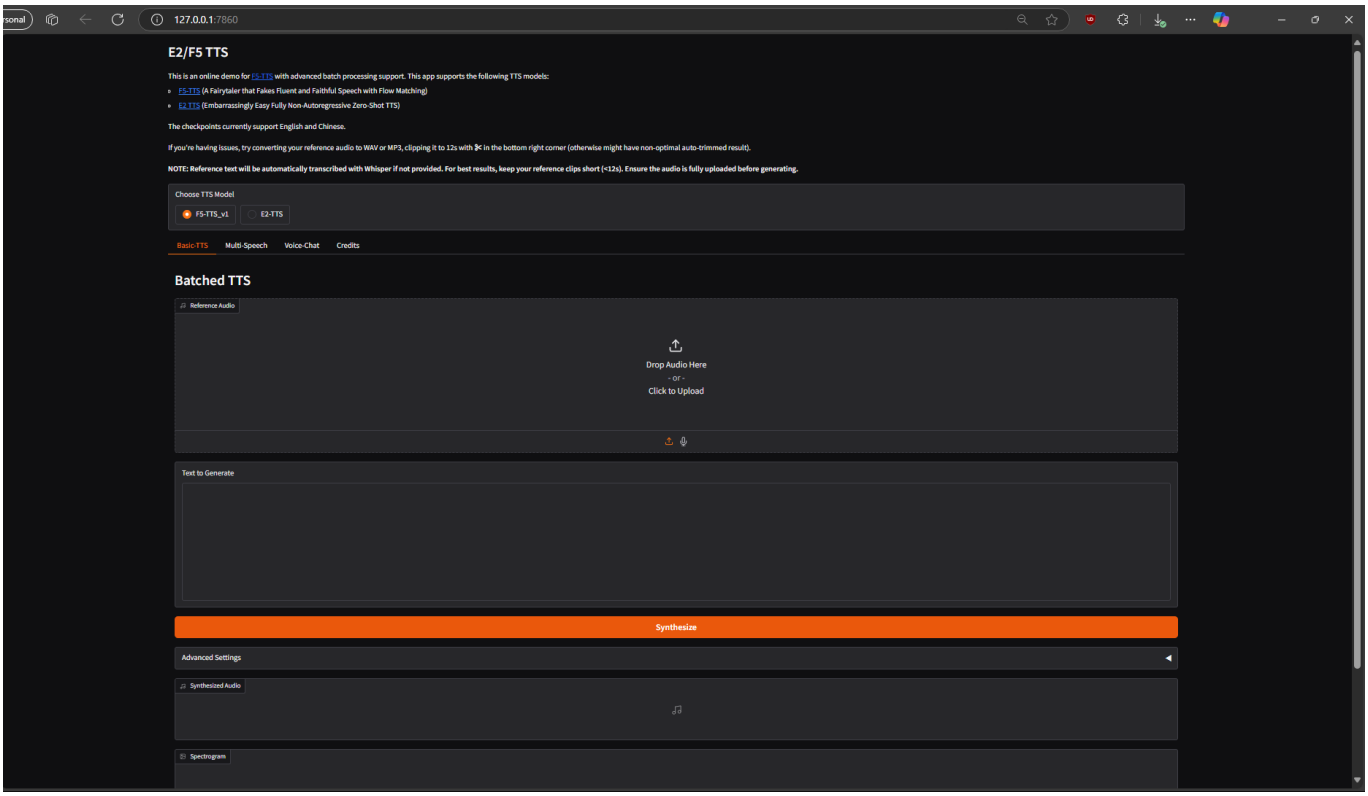
F5-TTS kết hợp các công nghệ tiên tiến để cải thiện hiệu suất và chất lượng tổng hợp giọng nói:

- **Diffusion Transformer với ConvNeXt V2**: Sự kết hợp này giúp tăng tốc độ huấn luyện và suy luận, đồng thời cải thiện chất lượng giọng nói tổng hợp.
- **E2 TTS (Flat-UNet Transformer)**: Đây là một mô hình được tái tạo gần nhất từ nghiên cứu gốc, giúp cải thiện độ chính xác và tự nhiên của giọng nói.
- **Sway Sampling**: Chiến lược lấy mẫu trong quá trình suy luận này giúp cải thiện đáng kể hiệu suất của hệ thống.

Quy trình tổng hợp giọng nói của F5-TTS bao gồm các bước chính:

1. **Tiền xử lý văn bản**: Chuyển đổi văn bản đầu vào thành dạng biểu diễn phù hợp.
2. **Mô hình hóa ngữ âm**: Sử dụng bộ mã hóa để biến văn bản thành các đặc trưng ngữ âm.
3. **Tổng hợp giọng nói**: Mô hình tạo ra tín hiệu giọng nói từ các đặc trưng ngữ âm.
4. **Hậu xử lý**: Chuẩn hóa âm thanh và xuất tệp giọng nói.

Demo



3) Phân tích

1. Điểm mạnh:

- Chất giọng và biểu cảm khá tốt
- File audio gen khá nhanh

2. Điểm yếu:

- Ít tủy chỉnh so với các model khác (xtts, MVC)
- Mô hình tiếng Việt thì có rồi, 1 model train 41h41h nhưng mà chưa public, 1 model khác là F5-TTS 100h Vietnamese rồi nhưng mà đang chờ authenticate
- Đối với mô hình tiếng Việt, chưa đọc được 1 số từ, không đọc được ngày/tháng/năm, cụ thể là các con số