Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
(повна назва)
Кафедра програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів генерації анімаційного зображення на основі аудіопотоку
(тема)

Виконав:
здобувачка 2 року навчання
групи ІПЗм-23-3
Олександра КОМІНА
(власне ім'я, прізвище)

Спеціальність 121 - Інженерія програмного забезпечення
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Керівник доц. Олексій ТУРУТА
(посада, власне ім'я, прізвище)

Допускається до захисту
Зав. кафедри _____ Кирило СМЕЛЯКОВ
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
Кафедра програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 - Інженерія програмного забезпечення
(код і повна назва)
Тип програми освітньо-наукова програма
(освітньо-професійна або освітньо-наукова)
Освітня програма Інженерія програмного забезпечення
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____20_ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Коміній Олександрі Олександрівні
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження методів генерації анімаційного зображення на основі аудіопотоку»
затверджена наказом університету від 15 квітня 2025 р. № 290Ст
2. Термін подання здобувачем роботи до екзаменаційної комісії 16 червня 2025р.
3. Вихідні дані роботи: календарний план роботи, методичні вказівки до оформлення пояснювальної записки, перелік методів навчання для класифікації текстових повідомлень
4. Перелік питань, що потрібно опрацювати у роботі: вступ, предметна галузь з оглядом існуючих підходів, їх обмежень та сучасних тенденцій, здійснити аналіз наукових і літературних джерел, чітко сформулювати задачу, провести теоретичне дослідження методів генерації анімаційного зображення, процесу встановлення і запуску та існуючих інструментів, виконати порівняльний аналіз підходів і інтерпретувати отримані результати..

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання	11.04.2025	Виконано
2	Аналіз предметної галузі і постановка задачі	15.04.2025	Виконано
3	Теоретичне дослідження	21.04.2025	Виконано
4	Практичне дослідження	01.05.2025	Виконано
5	Підготовка пояснювальної записки	10.05.2025	Виконано
6	Підготовка презентації та доповіді	12.05.2025	Виконано
7	Перевірка на плагіат	11.06.2025	Виконано
8	Нормоконтроль	11.06.2025	Виконано
9	Рецензування	12.06.2025	Виконано
10	Попередній захист	13.06.2025	Виконано
11	Занесення диплома в електронний архів	13.06.2025	Виконано
12	Допуск до захисту у зав. кафедри	16.06.2025	Виконано

Дата видачі завдання 11 квітня 2025р.

Здобувачка _____ Олександра КОМІНА
(підпис)

Керівник роботи _____ доц. Олексій ТУРУТА
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить: 63 с., 16 рис., 6 табл., 31 джерело.

МОВА ПРОГРАМУВАННЯ PYTHON, AI, ARTIFICIAL INTELLIGENCE, GENERATIVE AI, TALKING FACE, WINDOWS, VIRTUALBOX, UBUNTU

Об’єктом дослідження виступають моделі для генерації анімаційного зображення на основі аудіо-потоку, які також називають Talking Face.

Метою даної роботи є вивчення можливостей, які дають подібні моделі генерації, а також ознайомлення з основними технологіями, які використовуються в цій галузі ШІ та дослідження можливостей, представляємих даними програмами.

Методами дослідження є аналіз технічної документації обраних на дослідження моделей ШІ, створення програмної системи, в умовах якої буде відбуватись практичне тестування існуючих обраних для дослідження рішень на прикладі різних вихідних даних.

У результаті дослідження було визначено основні opensource-застосунки генеративного ШІ, проаналізовано предоставлену їх розробниками технічну документацію, визначено основні характеристики і можливу область для застосування. Окрім цього, було протестовано такі параметри, як легкість встановлення і використання даних застосунків, знайдено основні переваги і недоліки досліджуваних програм, протестовано якість генерації на різних вихідних зображеннях, після чого було ретельно проаналізовано результати і зроблено висновки.

Отримані в ході дослідження результати показують перспективність даного напряму для використання в різних областях, а також певні недоліки при різних форматах використання і при різних вихідних умовах на кшалт технічних можливостей, а також необхідність подальшого розвитку даних технологій, а також оптимізації і енергоефективності алгоритмів.

ARTIFICIAL INTELLIGENCE, AI, GENERATIVE AI, PYTHON PROGRAMMING LANGUAGE, TALKING FACE, WINDOWS, VIRTUALBOX, UBUNTU

The object of the study is models for generating an animated image taking into account the audio stream, which also use Talking Face.

The method of this work is the ability to study the data that similar generation models provide, as well as familiarization with the main technologies that have been used in this field of AI and research presented by possible data programs.

The research methods are the analysis of the technical documentation of the AI models selected for the study, the creation of a software system in which practical testing of the existing solutions selected for the study will take place on the example of various source data.

As a result of the study, the main opensource applications of generative AI were identified, the technical documentation provided by their developers was analyzed, the main characteristics and a possible area of application were determined. without this, such parameters as the ease of installation and use of these applications were tested, the main advantages and disadvantages of the studied programs were found, the quality of generation was tested on various source images, after which the results were finally analyzed and conclusions were drawn.

The results obtained during the study show the prospects of this area of use in various, as well as certain shortcomings in different formats of use and under different initial conditions in terms of technical capabilities, as well as the necessary further development of these technologies, as well as optimization and energy efficiency of algorithms.

Завідувачу кафедри
ПІ_____
(скорочена назва кафедри)
проф. Кирилу СМЕЛЯКОВУ_
(вчене звання, сласне ім'я, прізвище)

ЗАЯВА

щодо самостійності виконання кваліфікаційної роботи та можливості її публікації (та/або публікації анотації кваліфікаційної роботи) в електронному архіві відкритого доступу ElAr KhNURE

Я,_____Коміна Олександра Олександрівна_____
(прізвище, ім'я, по батькові)

здобувач вищої освіти на другому (магістерському) рівні вищої освіти академічної групи___ІПЗм-23-3_____

кафедра_____ програмної інженерії _____,
(повна назва кафедри)

заявляю: моя кваліфікаційна робота на тему «Дослідження методів генерації анімаційного зображення на основі аудіопотоку»_____,
(назва роботи)

що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в репозиторії "ElArKhNURE". Погоджуюся з авторським договором, відповідно до Положення про репозиторій ХНУРЕ "ElArKhNURE". Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений (а) з вимогами академічної доброчесності, згідно з якими виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

Дата Підпис

ЗМІСТ

ВСТУП

Розвиток ШІ трансформує наше звичне життя щодня і щомиті, створюючи нові можливості і нові нечувані раніше небезпеки, про які ми ще недавно і не думали. Одним з найактуальніших напрямів розвитку є генеративний штучний інтелект, також відомий як GenAI, який дозволяє генерувати нову інформацію, як то, наприклад, відео, аудіо, зображення, текст, після навчання на великому наборі навчальних даних такого формату.

Talking face – один з напрямів GenAI, назва якого походить від одноіменного типу відео, в яких увага глядача сконцентрована на обличчі говорячої людини в кадрі. Ці технології дають можливість «оживити» портрет, використовуючи натренований ШІ для анімації фото- або відео-референсу під надане користувачем аудіо.

Актуальність дослідження зумовлюється поширенням використання штучного інтелекту, а також розповсюдженістю його використання та появі його в щоденному житті кожного. Наприклад, технологія Talking face дає можливість оживити аудіо-матеріал за допомогою додавання до нього відеоряду, що відкриває безліч можливостей і користі при використанні в сфері освіти. Також ця технолія може стати в нагоді при анімуванні віртуальних помічників, наприклад віртуальних консультантів. Також анімовані штучним інтелектом обличчя можуть послугувати в анімуванні портретів вже мертвих видатних особистостей, а також навіть вигаданих персонажів з художніх творів, що може принести дуже велику користь і додати наочності для створенні тих же освітніх екскурсій, коли сам автор розповідатиме про експонати чи творчість автора, або це може робити його персонаж.

Метою роботи слугує мета дослідити основні існуючі open-source пропозиції для генерації Talking face, вивчити їх офіційну документацію, після теоретичного аналізу визначити основні риси і напрями застосування, зазначити практичну можливість використання застосунків з Windows та Linux на прикладі Ubuntu з використанням VirtualBox, провести аналіз легкості встановлення і застосування, орієнтуючись на приклад звичайного користувача ПК, а в кінці

проаналізувати результати генерації на різних прикладах зображень, слугуючих референсами-першоджерелами.

Очікуваним результатом слугуватиме розуміння технології Talking face і основних технологій, що використовуються в генерації відеоряду, а також розуміння принципу їх встановлення і застосування, аналіз основних переваг і недоліків різних застосунків, і наостанок, основні вимоги до вихідних зображень.

Очікуваний підсумок: виконання цього дослідження дозволить глибше зрозуміти можливості різних застосунків в якості генерації відеоряду з урахуванням аудіоряду.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Актуальність і обґрунтування дослідження

Генерація Talking Face – це одна з передових технологій, що відкриває безліч нових можливостей у багатьох галузях і областях за допомогою вміння створювати реалістичні анімовані обличчя, що синхронно вимовляють текст. Актуальність цієї теми обумовлюється зростанням попиту на персоналізацію контенту, іноваційні інструменти комунікації та автоматизацію рутинних завдань у сучасному світі.

В часи все більш активної діджиталізації і цифровізації контент мультимедіа стає все більш важливим рушієм бізнесу, освіти і розваг. Окрім цього, штучна генерація відео-ряду по аудіо дає можливість додати інклюзивний аспект, бо робить контент більш доступним для людей з вадами слуху.

Важливим аспектом, що потенційно може почати широко використовуватись в майбутньому, є дублювання медіа, таких як фільми, серіали, передачі, а можливо і звичайні відео в соцмережах, за допомогою цієї технології. Генеративний інтелект дасть можливість синхронізувати рух губ акторів зі звуками нової мови, на яку було перекладено аудіоряд, і які б інакше не співпадали. Це дозволить досягнути як інклюзивності, так і більшої якості адаптації контенту.

Окремою неприємною, але, на жаль, важливою актуальністю також є не найбільш порядне використання даних технологій – для створення відеофейків, що може досягти стратегічної важливості в області дезінформації і так званої інформаційної війни, що в умовах сьогодення є безперечно важливим для кожного. Однак навіть без воєнного використання ці технології можуть активно використовуватись шахраями для обдурювання як людей, так і технологічних систем.

Таким чином, це дослідження несе під собою безліч відкриваючихся перспектив та причин провести дослідження. З найбільш важливих варто зазначити потенціал для іновацій та академічну цінність дослідження. Оскільки цей напрям технологій з’явився нещодавно і існує дуже обмежений проміжок

часу, ця галузь генеративного штучного інтелекту є молодою і повною потенціалу. І через це існує багато досі не відкритих способів використання цих застосунків різними способами, і безліч можливих шляхів для розвитку цих технологій в майбутньому. Ця ситуацція також породжує бізнес-цінність, бо представники бізнесу шукають і знаходять для нових технологій практичне використання, наприклад, з метою автоматизації процесів, як в згаданій вище локалізації на різні мови. В ролі академічної цінності ж виступатиме саме розуміння принципів цієї технології, що дасть можливість глибже зрозуміти механізми роботи генерування відео-ряду за аудіо-референсом і дозволить розпізнавати сгенеровані відео-фейки. Або ж навпаки, створювати схоже з різною метою.

Також не варто забувати важливість соціальної складової, що дозволить застосувати інклюзивний підхід у більш широких колах.

1.2 Аналіз тенденцій та перспектив

Основними тенденціями сьогодення є неспадаючий інтерес суспільства і стрімка еволюція технологій генеративного ШІ. Як приклад, можна назвати удосконалення реалістичності зображення зображення і плавності порухів та зростання загальної якості генерування відеоряду, паралельно зі зменшенням візуальних артефактів генерації.

Внаслідок інтересу до галузі і пов’язаного з цим розвитку, можна помітити тенденцію введення в ужиток цих моделей ШІ-генерації в багатьох сферах та розширення місць її можливого застосування. Наприклад, моделі генерування відеоряду можна поєднувати і інтегрувати з іншими генеративними моделями ШІ, як то моделі, призначені для генерації аудіоряду.

Звичайно, окрім суспільного інтересу, варто також враховувати тенденцію обережностіі застереження відносно усього пов’язаного зі ШІ. Наприклад, можливість генерувати фейкові відео з високим реалізмом, в тому числі для маніпулювання чи шахрайських схем, викликатиме застереження і переживання щодо етичного використання даної технології і підіймає актуальність такої

важливої проблеми сьогодення, як потребу розробки політик і стандартів щодо використання генеративних моделей штучного інтелекту.

Проаналізувавши ці тенденції, можна визначити основну перспективність технології генеративного штучного інтелекту. Вивчення цієї галузі несе велику перспективу технологічних проривів через її іноваційність і відносно недовге існування з часу заснування, бо саме сфера штучного інтелекту наразі є основним джерелом нових відкриттів і інновацій. Окрім цього, можна очікувати подальше поширення технологій у різних галузях застосування, що, дуже ймовірно, приведе до послідуючого введення спеціалізованих норм та регуляцій, оскільки генеративний штучний інтелект займає все більш значуще місце в повсякденному житті і сучасній реальності.

1.3 Предметна галузь

Предметна галузь обраного напряму охоплює методи, технології та інструменти, що використовують для генерування анімованих облич, що мають можливість синхронно відтворювати текст, мовлення чи інші аудіовізуальні сигнали. Ця сфера міждисциплінарна, охоплюючи такі ключові напрями, як CV (Computer Vision – комп’ютерний зір), DL (Deep Learning – глибинне навчання), NLP (Natural language processing – обробка природньої мови) і технології мультимедіа. Комп’ютерний зір дає можливість розпізнавання облич, виокремлення, визначення і ідентифікації на ньому певних виразів і рис, а також дозволяє використовувати аналіз міміки. Глибинне навчання лежить в основі процесу генерації анімації, використовуючи великі масиви даних, опрацьовані під час тренування моделей. Обробка природної мови забезпечує аналіз аудіо-референсу, розпізнавання мовлення та синхронізацію відеоряду із звуковим супроводом.

Ключовими складовими цієї предметної області є алгоритми, моделі та джерела даних, які застосовуються як для тренування, так і для самої генерації. Особливе значення мають технології відстеження рухів, які забезпечують вищу точність як при аналізі, так і під час створення відео.

Технологія має широкий спектр застосування, серед основних напрямів – медіа, індустрія розваг, освітні проєкти та системи навчання, а також сфера бізнесу й маркетингу.

1.3.1 Сфери застосування

Технологія Talking Face знаходить широке застосування в різних галузях завдяки своїй здатності оживляти статичні зображення та синхронізувати рухи губ із аудіо.

Одним з напрямів застосування може бути галузь розваг. Ці технології могли б значно покращити дубляж фільмів, серіалів і мультфільмів. Також при необхідності можна за допомогою ШІ створити віртуальних ведучих для онлайн-каналів. Це вже використовується і зараз, але не має сильно великого поширення. Також можна оживити інтерактивних персонажів в іграх, що потенційно може допомогти зробити ігри реалістичніше і підвищити загальну якість виконання.

В сфері освіти можливо додати багато потенційних варіантів для покращення засвоєння матеріалу: віртуальні викладачі, які пояснюють матеріал «вголос», мовні курси з адаптацією відеоряду під різні мови, а також оживлення історичних осіб чи вигаданих персонажів для проведення уроку чи екскурсії.

В галузі маркетингу реклами відкривається можливість використовувати інтерактивні відео, сгенеровані ШІ, для безлічі цілей від реклами до презентації новинок на ринок. В якості облич можна використовувати як обличчя зірок, так і спеціально створених маскотів.

В сфері комунікацій вже зараз активно використовуються віртуальні аватари для відеоконференцій чи соціальних мереж, що передають міміку і інтонацію мовлення. З розвитком технології можлива поява такої технології, як автоматичний дубляж відеоряду, якщо комунікація відбувається між носіями різних мов.

В сфері інклюзивності основним акцентом було би створення відеоряду і субтитрів для людей з порушенням слуху. Також, можливо, створення

інтерактивних персонажів, які допомагають в навігації з застосунком, могло б допомогти навіть звичайним людям, які не вміють звертатись з технікою.

В галузі креативної індустрії ж можна відкрити такі напрями, як анімація художніх зображень, а також цифрове реставрування і оживлення архівних відеоматеріалів.

1.3.2 Мови програмування для Talking Face

Python[14] здобув домінуюче місце серед мов програмування для реалізації технологій Talking Face завдяки поєднанню багатьох переваг. По-перше, його екосистема надає вичерпний набір бібліотек для глибинного навчання, комп'ютерного зору, аудіоаналізу та обробки медіа. Інструменти на зразок PyTorch[9] або TensorFlow дозволяють будувати й навчати складні нейронні мережі для синхронізації рухів губ і генерації міміки, тоді як OpenCV і Dlib спрощують детекцію облич і відстеження ключових точок. Аудітехнічні модулі, як-от librosa або PyDub, забезпечують аналіз і перетворення звукових сигналів, необхідних для точного вирівнювання аудіо з відеорядом, а обгортки ffmpeg-python та MoviePy роблять обрізку, злиття й конвертацію медіафайлів доступними через декілька рядків коду.

У роботі над прототипами Talking Face важливим є швидкий цикл експериментів, і Python надає можливість миттєво перевіряти зміни без потреби у тривалому компілюванні. Інтерактивність Jupyter Notebook дозволяє у реальному часі візуалізувати графіки втрат, оцінювати результати синхронізації й оперативно коригувати архітектуру мереж. Хоч сама мова інтерпретована, обчислення у важких нейронних шарах виконуються у високооптимізованих C/C++/CUDA-бібліотеках, тому продуктивність виконання залишається високою, навіть при роботі з GPU.

Ще однією сильною стороною є «клейкість» Python-коду: модель глибинного навчання, медіапроцесинг, аудіоаналітика та навіть веб- або десктопний інтерфейс можуть бути органічно об'єднані в єдину програму. Завдяки активній спільноті та численним відкритим репозиторіям практично

кожна нова розробка з Talking Face — від Wav2Lip до MakeItTalk — доступна у вигляді готового Python-коду, що значно спрощує впровадження й налаштування.

Кросплатформеність Python гарантує однакову працездатність як на Windows, так і на Linux чи macOS, а менеджмент залежностей через pip або Conda і віртуальні середовища дозволяють уникнути конфліктів між версіями бібліотек. Нарешті, для виведення Talking Face-рішень у виробництво існують багаточисельні інструменти DevOps та MLOps – Docker, Kubernetes, MLflow, DVC тощо – які безшовно інтегруються з Python-проєктами, забезпечуючи надійність і масштабованість сервісів. Усе це робить Python природним вибором для дослідників та розробників, що створюють і впроваджують рішення зі «говорячими» обличчями.

C++ часто обирають для реалізації компонентів Talking Face там, де критично важлива швидкість обробки та низька затримка, адже ця мова компілюється в машинний код і дозволяє максимально ефективно використовувати ресурси процесора та відеокарти. Багато бібліотек комп’ютерного зору, зокрема OpenCV, написані на C++ з огляду на високу продуктивність, а бекенди таких фреймворків глибинного навчання, як TensorRT або ONNX Runtime, теж реалізовані переважно на C++ і надають відповідні API для виклику з інших мов. Крім того, C++ дає змогу створювати розширення та плагіни для Python-модулів, поєднуючи зручність прототипування з Python із необхідною швидкістю виконання в продакшені. У застосунках реального часу, наприклад для відеоконференцій із прогнозуванням міміки або для вбудованих систем на периферійних пристроях, C++ забезпечує детерміновану продуктивність і мінімальні накладні витрати на управління пам’яттю. Наявність стандартних бібліотек STL, можливість тонкого налаштування оптимізації компілятора і робота без непотрібних абстракцій роблять C++ незамінним у тих частинах Talking Face рішень, де будь-яка мілісекунда затримки може стати критичною.

2 ОГЛЯД Й АНАЛІЗ ЛІТЕРАТУРНИХ, НАУКОВИХ ДЖЕРЕЛ

2.1 Огляд основних існуючих моделей для генерації відеоряду

Цей розділ присвячено огляду існуючих моделей штучного інтелекту, що використовуються для генерації синхронізованого відеоряду, з коротким описом їх характеристик. Оскільки цей напрям дуже молодий, основні дослідження базуються саме на створенні нових систем, і таким чином, в цьому розділі ми проаналізуєм основні розроблені застосунки і їх документацію в якості наукових і літературних джерел.

Wav2Lip[2] – це сучасна модель штучного інтелекту, розроблена для створення реалістичної синхронізації рухів губ із аудіосигналом у відео. Її створили дослідники з Інституту інформаційних технологій в Хайдарабаді (IIIT-H). Модель дає змогу синхронізувати губи людини у відео з поданим звуковим супроводом, створюючи ефект, що людина справді вимовляє цей текст.

Для генерації відео необхідно мати відео із зображенням людини, текст, за яким анімується губна артикуляція, та аудіо, під яке здійснюється синхронізація відеоряду.

Існує дві версії моделі: Wav2Lip і Wav2Lip + GAN. Перша зосереджена насамперед на максимально точній синхронізації аудіо з відео. Друга, хоча й поступається в точності синхронізації, забезпечує вищу якість зображення завдяки використанню генеративно-змагальної мережі (GAN).

Для роботи моделі потрібні середовище Python 3.6 і наявність ffmpeg. При цьому система орієнтована на запуск у Unix-подібних операційних системах, що може спричинити значні труднощі при спробі інсталяції на Windows.

Розробники позиціонують модель як інструмент для синхронізації відеоряду при адаптації відео до інших мов і зазначають, що система підтримує всі мовні варіанти.

Модель є повністю безкоштовною для особистого або академічного використання.

DINet: Deformation Inpainting Network for Realistic Face Visually Dubbing on High Resolution Video[1] – це модель штучного інтелекту, призначена для

реалістичного дублювання рухів обличчя у відео високої роздільної здатності. Її головна задача – точно й правдоподібно замінювати частини обличчя, зокрема губи, для синхронізації з новим звуковим супроводом. На відміну від базових моделей, таких як Wav2Lip, DINet орієнтована на збереження високої якості та деталізації при роботі з відео у форматі високої чіткості.

Щоб згенерувати відео, необхідно мати оригінальне відео з людиною, текстовий супровід, за яким здійснюється анімація, та аудіофайл, з яким синхронізується відеоряд.

Для роботи моделі потрібні середовище Python 3.7 і ffmpeg. Водночас процес встановлення може ускладнюватися через необхідність використання специфічних версій бібліотек torch, torchvision і torchaudio, які наразі офіційно не підтримуються розробниками – тож їх доводиться завантажувати зі сторонніх джерел.

Попри це, DINet без проблем працює в операційній системі Windows, що робить її зручною та доступною для широкого кола користувачів завдяки універсальній процедурі інсталяції.

MakeItTalk[3] – це модель штучного інтелекту, яка трансформує статичне зображення обличчя у відео з анімованими рухами, синхронізованими з вхідним аудіо. Розроблена для створення реалістичних і переконливих говорячих персонажів, ця система дозволяє "оживляти" зображення, генеруючи природні рухи губ, щелепи та навіть міміки.

Однією з ключових особливостей моделі є те, що для генерації достатньо лише одного зображення – у тому числі намальованого або стилізованого персонажа. Це відкриває широкі можливості для візуальної анімації художніх творів, портретів історичних постатей чи випадків, коли є лише одне фото.

Для запуску системи необхідні Python 3.6, pynormalize[12] і ffmpeg. Водночас модель орієнтована переважно на Unix-подібні операційні системи, зокрема Ubuntu[17], на якій проводилося тестування. Поведінка моделі під час встановлення на Windows не документована, тому її стабільність на цій платформі невідома.

Оскільки модель працює лише з одним зображенням, рівень деталізації й реалістичності анімації поступається іншим рішенням. Тому вона не підходить для сфер, де критично важливий високий рівень достовірності – зокрема, у дубляжі фільмів, серіалів, створенні реалістичних відео або озвучуванні для людей із порушенням слуху.

Натомість, MakeItTalk ідеально підходить для освітніх або розважальних проєктів – наприклад, для анімації портретів відомих діячів чи оживлення намальованих персонажів, що відкриває творчі можливості у навчанні, культурі й медіа.

SadTalker[10] – це глибоконейронна модель, яка дозволяє генерувати анімоване відео говорячого обличчя на основі одного нерухомого зображення (портрету) та аудіофайлу (мовлення або музики). Модель використовує технологію 3DMM (3D Morphable Model), яка відтворює тривимірну структуру обличчя для забезпечення реалістичних рухів, зокрема губ, щелепи, очей і голови, синхронізованих із аудіо.

Для створення відео необхідно надати лише одне зображення, яке потрібно анімувати, та відповідний звуковий супровід. Завдяки цьому SadTalker може оживити як реальні фотографії, так і стилізовані зображення – наприклад, ілюстрації або портрети, створені художником.

Для роботи системи потрібна наявність ffmpeg, а також стандартне середовище Python. Особливістю моделі є її кросплатформеність – SadTalker можна запускати на Windows, Linux та macOS, що робить її доступною для широкого кола користувачів без значних технічних бар’єрів.

Оскільки модель працює лише з одним зображенням, вона не забезпечує ультрареалістичної якості відеоряду, тож її не варто застосовувати в професійному кіновиробництві, дубляжі чи спеціалізованих рішеннях для людей з порушенням слуху. Водночас SadTalker ідеально підходить для освітніх, культурних та розважальних проєктів – таких як анімація історичних персонажів, оживлення картин або створення інтерактивних персонажів для відеопрезентацій.

2.2 Огляд існуючих моделей для оцінювання якості штучно згенерованого відеоряду

Оцінка якості генерації моделей, таких як MakeItTalk, Wav2Lip та подібних, може здійснюватися за допомогою спеціалізованих інструментів та відкритих репозиторіїв. Вони дають змогу обчислювати показники якості синхронізації, природності та суб'єктивної (перцептивної) відповідності результатів. Далі розглянемо основні з них.

SyncNet[6] – це модель, створена для аналізу та покращення синхронізації між відео та аудіо, з особливим акцентом на відповідність рухів губ мовленню. Вона дозволяє оцінити точність синхронізації, вимірюючи часове відставання або випередження губ відносно аудіосигналу. SyncNet широко використовується для перевірки узгодженості між відеорядом і голосовим супроводом. Для запуску необхідні Python 3 і ffmpeg.

Q-Align[4] – це алгоритм, розроблений для точного узгодження рухів губ із аудіо, що особливо корисно у завданнях дубляжу, генерації анімованих персонажів або створення віртуальних співрозмовників. Окрім оцінки синхронізації, Q-Align здатен аналізувати якість зображення, що робить його більш універсальним у порівнянні з іншими інструментами. Водночас така багатофункціональність вимагає значних ресурсів пам’яті, що може бути обмеженням для менш потужних систем.

vBench[5] – це комплексний інструмент для тестування та оцінки моделей, які генерують анімовані відео, зокрема для перевірки синхронізації мовлення з відеорядом. Основне призначення vBench – створення єдиної стандартизованої платформи для об’єктивного порівняння якості різних моделей. Система дозволяє оцінювати відео за низкою критеріїв, таких як: Background Consistency, Subject Consistency, Overall Consistency, Temporal Style, Dynamic Degree, Aesthetic Quality, Temporal Flickering, Motion Smoothness, Imaging Quality тощо.

Існують дві версії інструменту: vBench для коротких відео (менше 5 секунд) і vBench-long для довших відео. Завантаження доступне через pip, однак наразі

інструмент не підтримується на Windows через відсутність однієї з необхідних бібліотек.

2.3 Огляд існуючих моделей відновлення облич

Моделі для відновлення обличчя – це спеціалізовані нейронні мережі або алгоритми, призначені для покращення якості та деталізації зображень облич. Їх основне завдання полягає у роботі з зображеннями низької якості, пошкодженими чи спотвореними – вони відновлюють відсутні елементи, зменшують рівень шуму, усувають артефакти й підвищують роздільну здатність.

Як вже зазначалося, такі моделі виконують низку важливих функцій, які будуть докладніше розглянуті далі. Зокрема, вони дозволяють покращити якість зображень з низькою роздільною здатністю, відновлюючи чіткість і реалістичність деталей, що особливо корисно при обробці старих фото або відеозаписів. Також моделі ефективно прибирають цифровий шум, зменшують артефакти стиснення JPEG чи інші дефекти, що погіршують якість візуального матеріалу. Крім цього, вони здатні відтворювати втрачені частини обличчя у випадках, коли фрагменти зображення розмиті, пошкоджені або зовсім відсутні. У цьому контексті подібні алгоритми є важливим доповненням до систем генерації відео, компенсуючи можливі візуальні похибки, створені під час автоматичної генерації.

VQFR (Vector Quantized Face Restoration)[7] – це сучасна модель для покращення зображень облич, яка використовує метод векторної квантизації для високоточного відновлення деталей. Її архітектура поєднує генеративні підходи з використанням попередньо навчених кодових книг, що дозволяє досягати високої реалістичності відновленого обличчя. Кожне зображення низької якості розбивається на частини, які зіставляються з відповідними векторами з кодової книги – ці коди представляють високоякісні текстури та деталі. Такий підхід дозволяє моделі враховувати як загальні риси обличчя (глобальний контекст), так і найдрібніші деталі (локальний контекст), що робить її ефективною навіть при роботі зі складними спотвореннями.

Для використання VQFR необхідні Python 3.7 або новіший, Pytorch версії 1.7 і вище, а також рекомендована робота в середовищі Linux. Аналогічні вимоги має і інша модель – GFPGAN.

GFPGAN (Generative Facial Prior GAN)[8] – ще одна передова модель для відновлення обличчя, створена з використанням генеративних змагальних мереж (GAN). Вона спрямована на відновлення реалістичних і деталізованих зображень із пошкоджених або низькоякісних джерел. Архітектура включає дві взаємодіючі нейронні мережі – генератор і дискримінатор: перший створює зображення, другий оцінює їхню якість, змушуючи генератор постійно вдосконалювати результати. Модель використовує попередні знання про анатомічну структуру людського обличчя, що дозволяє точніше відновлювати основні риси.

GFPGAN здатна відтворювати навіть найменші деталі – структуру шкіри, форму очей, лінії волосся та загальні риси обличчя – забезпечуючи максимально природний вигляд зображення. Як і VQFR, ця модель демонструє високу ефективність у застосуванні до генеративних систем для покращення візуальної якості кінцевого відеоряду.

2.4 Огляд додаткових необхідних програм

OpenFace[19] – це програмне забезпечення з відкритим кодом, розроблене для розпізнавання та аналізу облич. Воно дозволяє знаходити обличчя на фото та у відео, визначати ключові точки обличчя (facial landmarks), розпізнавати емоційний стан, орієнтацію голови, напрямок погляду, а також ідентифікувати особу.

Основні функції OpenFace включають:

- оцінка напрямку погляду (Gaze Estimation) – обраховує, куди саме спрямований погляд користувача;
- визначення ключових точок обличчя (Facial Landmark Detection) – виявляє до 68 характерних точок, серед яких очі, брови, рот, ніс і контур обличчя;

- визначення м'язових скорочень (Facial Action Unit Detection) – виявляє активність м’язів обличчя відповідно до шкали FACS, що використовується для аналізу емоцій;
- розпізнавання обличчя (Face Recognition) – використовується для ідентифікації особистості на основі зображення обличчя.
- оцінка положення голови (Head Pose Estimation) – визначає орієнтацію голови в тривимірному просторі (yaw, pitch, roll).

FFmpeg[16] – потужна кросплатформна утиліта з відкритим кодом для обробки відео, аудіо та потокового контенту. Це універсальний інструмент, який дозволяє виконувати широкий спектр операцій із медіа: конвертація форматів, вирізання або поєднання фрагментів, запис, зміна частоти кадрів, швидкості, накладання звуку, субтитрів, логотипів, фільтрів тощо.

FFmpeg підтримує практично всі популярні формати (.mp4, .avi, .webm, .mkv, .mp3, .aac, .flac, .gif і багато інших). Він не потребує графічного інтерфейсу – всі дії виконуються через командний рядок, що дозволяє легко автоматизувати процеси обробки медіа через скрипти. Завдяки своїй швидкості, гнучкості та широкому функціоналу FFmpeg є ключовим інструментом при створенні штучно згенерованих відео, зокрема як обов'язковий компонент для багатьох моделей генерації.

Pip – це основний інструмент для встановлення та управління Python-бібліотеками з репозиторію PyPI. З його допомогою можна легко інсталювати, оновлювати або видаляти пакети, переглядати встановлені бібліотеки, зберігати перелік залежностей у файл (requirements.txt) тощо. Завдяки pip процес інсталяції необхідних для роботи застосунків бібліотек є швидким і зручним.

Virtualenv[15] – це утиліта для створення ізольованих середовищ Python, що дає змогу уникнути конфліктів між версіями пакетів у різних проєктах. Кожне середовище створюється окремо і містить власний інтерпретатор Python та набір бібліотек. Це особливо корисно, коли для різних проєктів потрібні різні версії одних і тих самих пакетів або самого Python.

Git[13] – система контролю версій, яка забезпечує відстеження змін у коді, збереження історії розробки, роботу з різними версіями (гілками) проєкту та співпрацю в команді. Git дозволяє зберігати резервні копії, відновлювати попередні стани, зручно інтегрується з сервісами на кшталт GitHub, GitLab або Bitbucket. У контексті цієї роботи Git використовується для встановлення необхідних застосунків безпосередньо з репозиторіїв через командний рядок.

2.5 Узагальнення результатів огляду джерел

Огляд літератури та аналіз наявних інструментів чітко демонструють, що Talking face генерація займає важливе місце серед інших технологій штучного інтелекту, і має свої переваги і сфери застосування. Однак, оскільки на даний момент ця галузь, як і галузь генеративного штучного інтелекту, ще нова і недосліджена, існує доволі мало матеріалу по дослідженням. Також варто зазначити, що всі дослідження і розробки концентруються на англійській мові, як і датасети для навчання ШІ. Так чи інакше, ця технологія продовжує розвиватись, що підтверджує необхідність досліджень цієї галузі. Також варто зазначити, що в дослідженнях зазвичай не підіймається тема зручності користування на користувацьких ПК, які не мають потужності промислових серверів, на які орієнтується розробка ШІ. Таким чином ми можемо підтвердити актуальність дослідження обраної теми, бо вона актуальна в умовах сьогодення, вона при цьому відносно неглибоко вивчена, і мало досліджень з точки зору звичайного користувача. Ми змогли виділити в ході дослідження літератури і наукових джерел основні програми для генерації, а також основні допоміжні програми в галузі, як то оцінювачі якості сгенерованого відео, застосунки відновлення якості фотографій і застосунки для відслідковування виразів обличчя. Основна цінність дослідження полягає в вивченні цих існуючих застосунків і їх практичної цінності для використання на стандартному ПК.

У перспективі майбутні дослідження та розробки, найімовірніше, зосередяться на збільшенні якості і реалістичності генерації, можливо, також покращивши їх енергоефективність і швидкість генерації.

3 ПОСТАНОВКА ЗАДАЧІ

3.1 Опис предметної області та мети дослідження

Метою даної роботи є аналіз основних open-source рішень для генерації Talking face, ознайомлення з їхньою документацією, виявлення ключових характеристик та можливих сфер застосування. Подальшим етапом є практична перевірка можливості використання цих інструментів звичайними користувачами ПК, оцінка зручності встановлення та експлуатації, а також порівняння на основі визначених параметрів.

Окрему увагу приділено дослідженню якості роботи програм при використанні різних типів вхідних даних, зокрема у складних випадках (наприклад, із затемненими чи частково закритими обличчями), з метою виявлення обмежень відповідних моделей.

На основі вищезазначеного сформульовано такі основні завдання:

- провести аналіз основних open-source рішень для генерації Talking face;
- ознайомитись із документацією до вибраних рішень, визначити їхні ключові можливості та напрями застосування;
- відібрати найбільш придатні моделі генерації відео на основі документації та протестувати їх на ОС Windows або Ubuntu (у середовищі VirtualBox);
- оцінити успішність встановлення моделей;
- дослідити продуктивність моделей на складних вхідних даних та сформулювати на основі цього відповідні обмеження.

За результатами дослідження будуть сформульовані практичні рекомендації щодо вибору оптимальних моделей з урахуванням операційної системи та інших умов застосування. Це дозволить оцінити доцільність використання відповідних рішень у реальних сценаріях.

3.2 Обґрунтування вибору методів дослідження

У цьому розділі обґрунтовано вибір дослідницьких методів, які використовуються для досягнення поставленої мети та реалізації визначених

завдань, пов’язаних із вивченням можливості застосування технологій генерації Talking face відеоряду на ОС Windows та Ubuntu у віртуальному середовищі VirtualBox[18].

Методологія дослідження побудована на комплексному підході, який поєднує теоретичні та практичні методи, адаптовані до специфіки теми.

Теоретичні методи:

– аналіз літератури та документації – передбачає вивчення наукових джерел, технічної документації, специфікацій та інших матеріалів, що стосуються моделей генерації Talking face. Це дозволяє сформувати цілісне уявлення про предмет дослідження, виявити актуальні проблеми та сформулювати дослідницькі задачі;
– порівняльний аналіз – використовується для порівняння обраних рішень за основними характеристиками, можливостями, вимогами до середовища, що дозволяє виявити їх переваги, недоліки та області ефективного використання.

Практичні методи:

– розробка прототипу середовища – одним із центральних елементів цього дослідження є створення програмної системи, яка буде здатна використовувати застосунки, на яких проводиться дослідження. Метод передбачає створення тестового середовища на базі ОС Ubuntu у VirtualBox, у якому проводиться встановлення та перевірка роботи моделей. Це дозволяє практично оцінити легкість інсталяції, вимоги до ресурсів, зручність взаємодії з інтерфейсом, а також перевірити працездатність моделей на різних вхідних даних. Робота з прототипом дозволяє оцінити такі критерії, як легкість встановлення, легкість використання, використовуєма пам’ять, а також дозволяє застосувати встановлені застосунки на різних вихідних зображеннях і проаналізувати результати. Цей процес, окрім закладання фундаменту для послідуючих досліджень, також сприяє виявленню технічних труднощів, які можуть

виникнути при інтеграції досліджуваних застосунків з системами користувачів, і формуванню стратегій для їх подолання;

- експериментальне дослідження – реалізується шляхом тестування відібраних моделей на різних типах ОС (Windows, Ubuntu) та з різними вхідними зображеннями, що дозволяє оцінити стабільність, якість результатів, а також здатність програм коректно працювати в різних умовах;
- аналіз результатів – обробка та інтерпретація експериментальних даних із використанням візуалізацій (таблиць, графіків) з метою отримання кількісної та якісної оцінки ефективності моделей. Це дозволить зробити висновки щодо зручності застосування, якості генерованого відео та витрат ресурсів, а також визначити обмеження та потенційні напрями вдосконалення технологій;
- опитування та експертна оцінка (за наявності) – передбачає залучення думок користувачів або фахівців для уточнення переваг та недоліків застосунків. За відсутності можливості проведення опитування акцент робиться на об’єктивному аналізі експериментальних даних.

Обрані методи дослідження є достатніми та релевантними для досягнення поставленої мети. Теоретичний аналіз дає змогу глибше зрозуміти предметну область, тоді як практична перевірка забезпечує об’єктивну оцінку ефективності програмних рішень у реальних умовах. Такий інтегрований підхід гарантує всебічне та обґрунтоване дослідження теми.

4 ТЕОРЕТИЧНЕ ДОСЛІДЖЕННЯ

На основі аналізу документації було опрацьовано п’ять моделей генерації синхронізованого відеоряду, обраних для подальшого дослідження.

Wav2Lip – це модель штучного інтелекту, яка забезпечує реалістичну синхронізацію рухів губ із вхідним аудіо. Її головне призначення – створювати ефект, ніби людина на відео дійсно озвучує поданий звуковий фрагмент. Для роботи моделі необхідне відео з обличчям людини, аудіо, з яким потрібно синхронізувати губи, та, зазвичай, текст, який передає зміст. У рамках дослідження, оскільки аналізується саме генерація зображення, а не повноцінне відео, вихідне зображення трансформується у короткий відеофрагмент, створений з одного кадру, аби забезпечити рівні умови тестування для всіх моделей.

Цей застосунок включає дві навчені моделі: Wav2Lip, яка демонструє кращу точність у синхронізації, та Wav2Lip-gan, яка, використовуючи технологію GAN, забезпечує вищу якість зображення, хоч і з дещо гіршою синхронністю.

Wav2Lip-HD[11] є покращеною версією Wav2Lip, з додатковим використанням ESRGAN для підвищення чіткості зображення.

DINet – ще одна модель, орієнтована на відтворення міміки у відео високої роздільної здатності. На відміну від Wav2Lip, вона додатково потребує файл із даними про розташування ключових точок обличчя (face landmarks), згенерований за допомогою інструменту OpenFace. В іншому підхід до генерації та тестування був аналогічним, із використанням відео, створеного з одного зображення.

MakeItTalk – це модель, що дозволяє створити відеоанімацію з одного зображення, синхронізовану з аудіо. Вона забезпечує базову міміку та синхронність, але не передбачає високої деталізації.

SadTalker також працює з одним зображенням, зокрема й зі стилізованими чи намальованими персонажами. Вона використовує GFPGAN для покращення якості обличчя, однак створює відео меншої якості порівняно з попередніми моделями. До того ж, офіційна підтримка SadTalker зосереджена переважно на Unix-подібних операційних системах, що ускладнює її тестування у вибраних середовищах.

5 ПРАКТИЧНЕ ДОСЛІДЖЕННЯ

5.1 Прототип

У межах дослідження було прийнято рішення працювати з двома операційними системами: Ubuntu 16.04 та Windows 10. Вибір обумовлений прагненням моделювати умови, наближені до користування звичайним користувачем, без залучення спеціалізованого обладнання чи складних конфігурацій. Windows 10 – одна з найпоширеніших операційних систем серед масового користувача, тоді як більшість моделей штучного інтелекту, які розглядаються в дослідженні, орієнтовані переважно на Unix-подібні системи. Зважаючи на це, для забезпечення належного тестування було обрано запускати Ubuntu 16.04 у віртуальному середовищі за допомогою VirtualBox.

Обидві системи працювали на одному комп’ютері з 16 ГБ оперативної пам’яті, що відповідає середнім технічним характеристикам типового персонального комп’ютера. Такий підхід дає змогу адекватно оцінити продуктивність застосунків в умовах, наближених до реальних (див.рис.5.1).

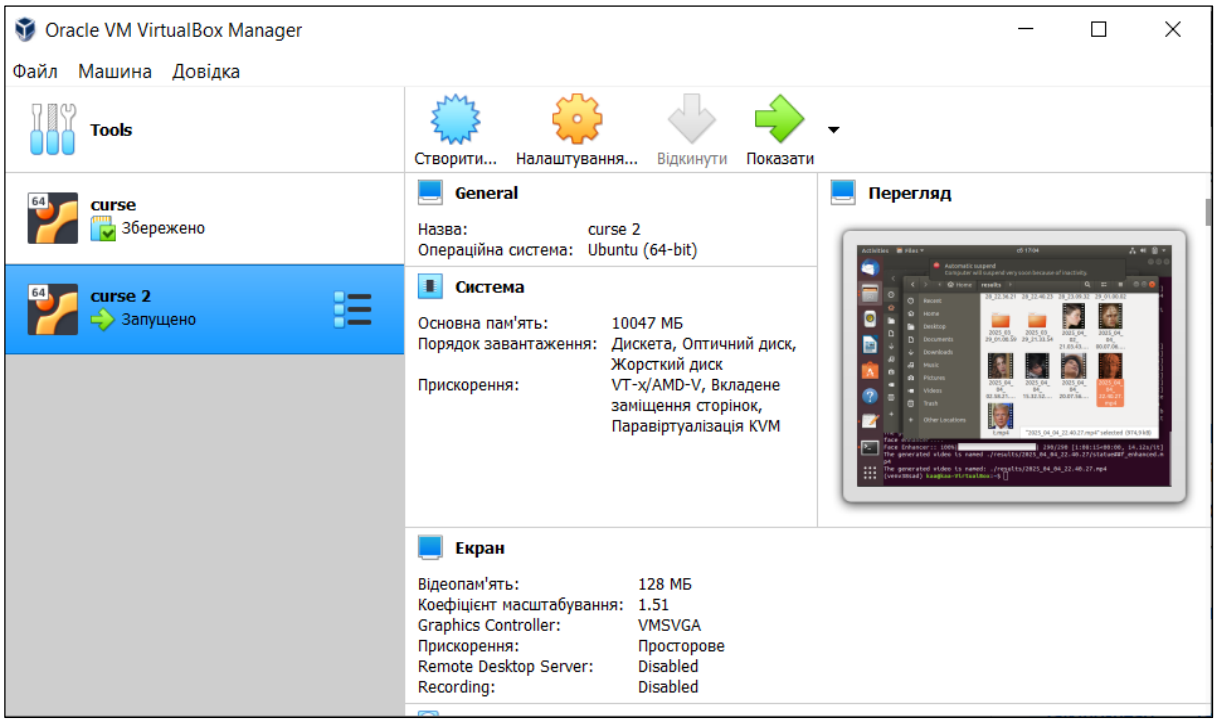

Рисунок 5.1 – Створена віртуальна машина (рисунок створено самостійно)

Для коректного розгортання тестованих моделей та забезпечення їх повноцінної роботи було встановлено низку допоміжних програм на обидві системи. Зокрема, FFmpeg застосовувався для роботи з медіафайлами (зображеннями, аудіо та відео) і був необхідний як на етапі підготовки відео з одного зображення, так і для подальшої обробки відеоматеріалу. Git використовувався для швидкого та зручного завантаження вихідного коду застосунків з відкритих репозиторіїв. Pip служив основним менеджером пакетів для інсталяції необхідних бібліотек. Для ізоляції середовищ запуску кожної моделі, запобігання конфліктам між різними версіями залежностей та підтримання стабільної роботи всіх компонентів, застосовувалася утиліта Virtualenv.

Після завершення теоретичної частини дослідження та створення прототипу було проведено практичне тестування, під час якого здійснено спроби встановлення та використання згаданих моделей у середовищі операційної системи Windows та на віртуальній машині з Ubuntu.

5.2 Практичне дослідження на легкість встановлення

У ході перевірки з’ясувалося, що модель Wav2Lip не встановлюється на Windows – процес установки застрягає в нескінченному циклі, який завершується помилкою. Це дозволяє зробити висновок, що дана модель непридатна для генерації відеоряду в середовищі Windows. Відповідно, будь-які модифікації чи версії, побудовані на основі Wav2Lip, також не можуть бути застосовані в цьому середовищі. Попри це, модель заявлена розробниками як сумісна з Windows, що дає підстави припустити, що виявлена помилка могла бути зумовлена специфікою тестової конфігурації. Проте, навіть попри спроби усунути проблему протягом прийнятного часу, її не вдалося вирішити. Це дозволяє дійти висновку, що використання Wav2Lip на Windows пов’язане з істотними труднощами.

Натомість під час тестування на Ubuntu встановлення пройшло успішно, хоча й потребувало вирішення кількох незначних технічних нюансів (див.рис.5.2).

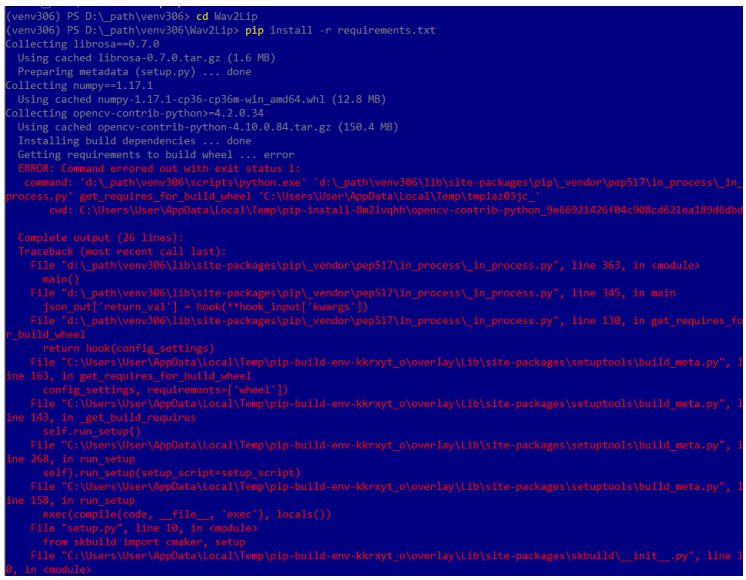

Рисунок 5.2 – Результати спроби встановлення Wav2Lip на Windows (рисунок створено самостійно)

Після чого було проведено спробу встановити DINet, що відбулось успішно. Після чого успішність встановлення була протестована через запуск генерації за датасетами від розробників (див.рис.5.3).

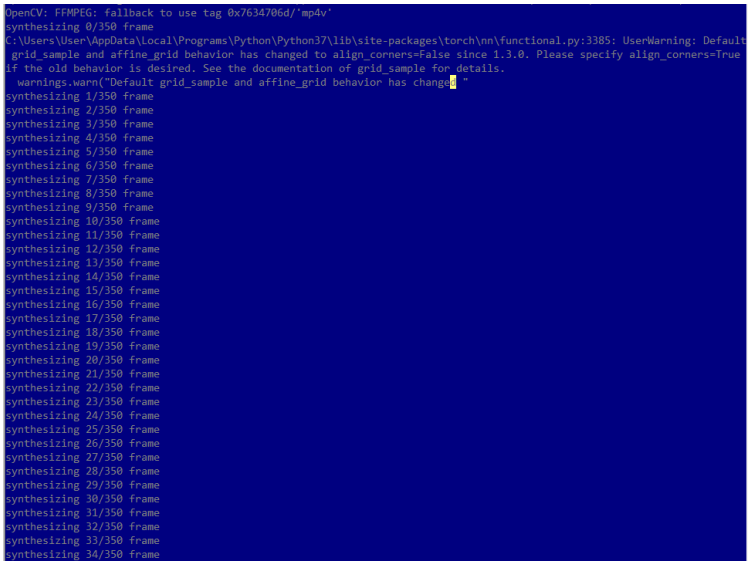

Рисунок 5.3 – Процес генерації відеоряду в DINet (рисунок створено самостійно)

Як результат було отримано два відео, одне з яких – фінальний результат, інше – сгенерований фрагмент обличчя, рухаючийся відповідно до вихідного аудіо (див.рис.5.4).

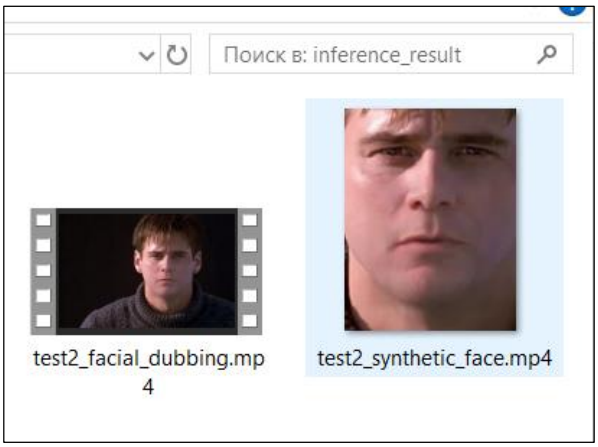

Рисунок 5.4 – Результат генерації відеоряду в DINet (рисунок створено самостійно)

Проте слід зауважити, що для повноцінного використання моделі DINet з власними зображеннями, а не лише з дефолтними, необхідно окремо встановити OpenFace, що вимагає додаткових налаштувань і займає чимало місця на диску. Крім того, оскільки DINet, на відміну від інших розглянутих моделей, написано мовою програмування C++, її встановлення на Linux-систему супроводжується певними труднощами. Натомість запуск цієї моделі на Windows відбувається без ускладнень, що контрастує з ситуацією з Wav2Lip, яка успішно працює саме на Unix-подібних системах.

Далі було перевірено роботу MakeItTalk на обох платформах. Попри відносно просте встановлення, виникла проблема – після перевірки наявності CUDA модель припиняє виконання, якщо графічний процесор недоступний. Це свідчить про те, що успішне використання деяких моделей може критично залежати від апаратної конфігурації комп’ютера. Як наслідок, користувачі з менш потужними ПК можуть зіткнутись із труднощами, що варто враховувати під час вибору технології для практичного застосування.

Наступною була протестована модель SadTalker, яка встановилась без жодних проблем як на Ubuntu, так і на Windows.

Окремо варто згадати інструменти для оцінки якості генерації, такі як vbench та Q-Align. Запуск обох виявився неможливим через обмеження ресурсів комп’ютера. У випадку з Q-Align на GitHub зазначалося, що проблему можна

вирішити, збільшивши обсяг оперативної пам’яті з 64 до 96 ГБ. Натомість vbench після запуску просто зависав у процесі завантаження. Це дозволяє зробити висновок, що автоматизована оцінка якості генерації виявилась недоступною через недостатні технічні характеристики пристрою, на якому проводилось тестування.

5.3 Аналіз результатів виконання програм

Першим етапом практичного тестування було запуск моделі Wav2Lip. Для цього ми використали звичайне зображення у фронтальному ракурсі, обрізане до рекомендованих розмірів (див.рис.5.5).

Рисунок 5.5 – Вихідне зображення (рисунок створено самостійно)

В результаті отримуєм відео, згенероване за декілька хвилин (див.рис.5.6)

Рисунок 5.6 – Кадр зі згенерованого Wav2Lip відео (рисунок створено самостійно)

На вихідному відео спостерігається розмитість зображення, особливо помітна в зоні рота, яка виглядає як чітко окреслений квадрат. Водночас, незважаючи на розмитість, рухи губ виглядають досить реалістично і плавно.

Далі було здійснено спробу запуску версії Wav2Lip-gan, однак процес було перервано через помилку. Це дало підстави для висновку, що версії Wav2Lip-gan та Wav2Lip-HD (яка також базується на Wav2Lip-gan) мають значно вищі вимоги до обсягу оперативної пам’яті, що ускладнює їх використання для пересічного користувача (див.рис.5.7).

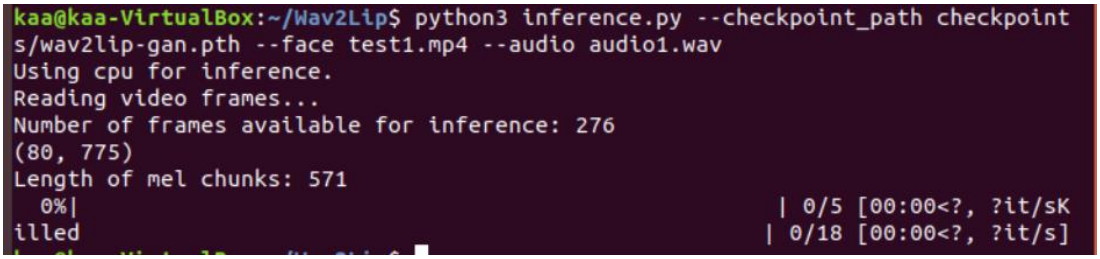

Рисунок 5.7 – Помилка через нестачу оперативної пам’яті при генерації з Wav2Lip-gan (рисунок створено самостійно)

Наступним етапом стало тестування SadTalker. Цей застосунок працює повільно – на генерацію 11 секунд відео пішло понад три години. Однак, отриманий результат значно перевершує інші протестовані моделі за якістю. Важливо зазначити, що SadTalker, на відміну від інших рішень, додає не лише рухи рота, а й голови, що робить відео більш динамічним і живим. Проте це також може порушити атмосферу оригінального зображення, тому така поведінка моделі має враховуватись при виборі інструменту (див.рис.5.8).

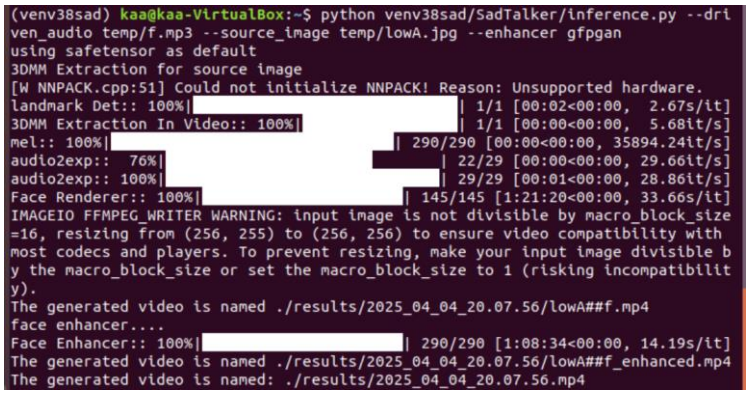

Рисунок 5.8 – Приклад виконання генерації в SadTalker (рисунок створено самостійно)

Важливо зазначити, що SadTalker, на відміну від інших рішень, додає не лише рухи рота, а й голови, що робить відео більш динамічним і живим. Проте це також може порушити атмосферу оригінального зображення, тому така поведінка моделі має враховуватись при виборі інструменту (див.рис.5.9).

Рисунок 5.9 – Кадр з результату генерації в SadTalker (рисунок створено самостійно)

Далі було протестовано DINet. Для запуску на власних зображеннях необхідно попередньо обробити фото за допомогою OpenFace, згенерувавши файл з параметрами, такими як координати facial landmarks. Попри виконання цього кроку, модель відмовилась працювати з отриманими даними через невідому помилку, що демонструє складність і нестабільність використання DINet. Тим не менш, на прикладі вбудованих демонстрацій видно, що ця модель генерує якісні результати, подібно до Wav2Lip, але з кращим переходом в області рота – межі квадрату практично непомітні. За швидкістю виконання DINet повільніша за Wav2Lip, але помітно швидша за SadTalker (див.рис.5.10).

Рисунок 5.10 – Кадр з результату генерації в DINet (рисунок створено самостійно)

5.4 Дослідження на різних вихідних зображеннях

Для глибшого аналізу була обрана модель SadTalker – як така, що демонструє найвищу якість генерації. Метою стало вивчення, як вона справляється зі складними умовами вхідних зображень. Спершу було протестовано генерацію відео на основі зображення в профіль. Згенероване відео виглядає цілком реалістично, на рівні з результатами для фронтальних зображень. Водночас спостерігаються певні недоліки: викривлення пропорцій обличчя – зміна форми носа, вух, загального співвідношення частин обличчя. Також текстура шкіри стала менш деталізованою, а тінь під підборіддям була інтерпретована як розмита пляма, що знижує реалістичність відео (див.рис.5.11).

Рисунок 5.11 – Вихідне зображення і кадр з результату генерації по фото профіля (рисунок створено самостійно)

Наступним кроком було тестування зображень, зроблених з високого та низького кутів. У випадку високого кута, окрім уже зазначених втрат текстури, виявлено проблему з другим оком, частково прихованим на фото. У відео воно або деформується, або зникає. Очі також зазнають змін: модель змінює їхній колір і напрям погляду відповідно до положення голови, а не першоджерела. Також ШІ змінює характер освітлення, роблячи його м’якшим і менш контрастним (див.рис.5.12).

Рисунок 5.12 – Вихідне зображення і кадр з результату генерації по фото з високого кута (рисунок створено самостійно)

У випадку з низьким кутом результати були найгіршими: викривлення обличчя стали найпомітнішими, а текстура волосся також зазнала значних спотворень, перетворившись на нечітку масу (див.рис.5.13).

Рисунок 5.13 – Вихідне зображення і кадр з результату генерації по фото з низького кута (рисунок створено самостійно)

Далі було досліджено роботу моделі з зображеннями, що мають яскраве та нестандартне освітлення. Як і в попередніх випадках, освітлення було згладжено, а світлі плями інтерпретовано як частину шкіри. Через це при русі голови вони залишаються на місці. Також ШІ не зберіг деякі важливі деталі, зокрема родимки та прядки волосся: вони були або стерті, або спотворені – прядка стала схожа на шрам, інша виглядала як затемнення на обличчі (див.рис.5.14).

Рисунок 5.14 – Вихідне зображення і кадр з результату генерації по фото з яскравим освітленням (рисунок створено самостійно)

Наступним було перевірено обробку зображення з елементами, що частково закривають обличчя. Як і раніше, втрачено текстуру шкіри, а також веснянки – ключову рису обличчя, що впливає на впізнаваність. Об’єкти на передньому плані, як-от рама чи пальці, сильно деформувались під час руху голови. Частина обличчя, яка була частково закрита, також згенерована з помилками, а сережка змінила форму й втратила свою геометричну точність (див.рис.5.15).

Рисунок 5.15 – Вихідне зображення і кадр з результату генерації по фото з частково закритим обличчям

На завершення було досліджено роботу SadTalker із штучним зображенням. Для цього використано фото голови статуї як приклад максимально реалістичного, але не живого образу. Найбільш очевидною зміною стало додавання очей. Це свідчить про тенденцію моделі надавати стилізованим об’єктам людських рис, що може як покращити загальне сприйняття, так і створити ефект «зловісної долини». Відповідно, це потрібно враховувати під час створення відео на основі стилізованих зображень (див.рис.5.16).

Рисунок 5.16 – Вихідне зображення і кадр з результату генерації по фото статуї (рисунок створено самостійно)

5.5 Аналіз і порівняння досліджених застосунків

На цьому етапі нами було сформульовано змістовну постановку багатокритеріальної задачі прийняття рішень для сфери ШІ-генерації анімаційного зображення на основі аудіопотоку. Після цього етапу буде проведено інформаційну підготовку прийняття рішення, а також буде сформульовано векторний опис задачі та за допомогою згорткової моделі теорії корисності буде знайдено її найкраще рішення.

В якості альтернатив буде проаналізовано Wav2Lip, Wav2Lip-GAN, Wav2Lip-HD, DINet, MakeItTalk та SadTalker.

Критерії для оцінки моделей генерації анімаційного зображення на основі аудіо-потоку:

а) якість і ефективність: відображає точність та ефективність генерації інструментом відео та синхронізації з аудіо:

атегоріальна шкала: Низька, середня, висока;

ислова шкала: оцінка ефективності від 1 до 5, де 1 – дуже низька якість, а 5 – висока;

б) доступність і легкість використання: відображає, наскільки просто встановити інструмент і працювати з ним, а також зручність його налаштування та використання для розробників:

атегоріальна шкала: дуже не зручно, не зручно, нормально, зручно, дуже зручно;

ислова шкала: від 1 до 5, де 1 - дуже не зручно, 5 - дуже зручно;

в) часова ефективність: відображає швидкість генерації відео:

атегоріальна шкала: менше хвилини, менше 10 хвилин, менше півгодини, менше години, година і більше (в контекті одного і того ж аудіоряду з фіксованою довжиною на одному і тому ж ПК);

ислова шкала: Від 1 до 5, де 5 – менше хвилини (дуже швидко), 1 – година і більше (дуже повільно);

г) вимогливість до ресурсів ПК: визначає рівень вимог для встановлення, запуску і генерації відео, наприклад, кількість оперативної пам'яті і необхідність графічного процесора; можливість запуску на слабкому ПК:

атегоріальна шкала: малі вимоги, середні, високі;

ислова шкала: від 1 до 5, де 1 – мало вимог, 5 – дуже високі вимоги;

д) об'єм займаємої пам'яті: відображає вимоги до пам'яті, яка необхідна для завантаження необхідних пакетів і моделей:

атегоріальна шкала: мало пам'яті (3 ГБ і менше), середньо пам'яті (5 ГБ і менше), багато пам'яті (більше 7 ГБ);

ислова шкала: від 1 до 5, де 1 – менше 3 ГБ, 5 – більше 7.

е) багатоплатформеність: оцінює, наскільки інструмент протестовано і розраховано і розроблено для високистовування на різних операційних системах:

атегоріальна шкала: нема багатоплатформенності, дві ОС, три ОС;

ислова шкала: від 1 до 3, де число – кількість з основних ОС.

Усі значення будуть виставлені за експертною оцінкою.

Векторний опис задачі багатокритеріального прийняття рішень для моделей генерування відеоряду по заданому аудіо:

Нехай X=(f1,f2,f3,f4,f5,f6) – альтернативи моделей генерування (формула

$$Xn = (f1 + f2 + f3 + f4 + f5 + f6)$$

де f1 – якість і ефективність,

– доступність і легкість використання,

– часова ефективність,

– вимогливість до ресурсів ПК,

– об'єм займаємої пам'яті,

– багатоплатформеність

Кожен компонент вектору представляє рівень кожного критерію для кожної альтернативи моделей генерації відеоряду.

Проводимо лінійну адитивну згортку з ваговими коефіцієнтами. Створимо табличне надання (табл.5.1)

Таблиця 5.1 – Лінійна адитивна згортка з ваговими коефіцієнтами (таблиця створена самостійно)

	Якість і ефективність	Доступність і легкість використання	Часова ефективність	Вимогливість до ресурсів ПК	Об'єм займаємої пам'яті	Багатоплатформеність
Wav2Lip	Низька	Нормально	Менше 10 хвилин	Середні вимоги	3,8 ГБ	Три ОС
Wav2Lip - GAN	Достатня	Нормально	Менше 10	Високі вимоги	4 ГБ	Три ОС

Кінець таблиці 5.1

	Якість і ефективність	Доступність і легкість використання	Часова ефективність	Вимогливість до ресурсів ПК	Об'єм займаємої пам'яті	Багатоплатформеність
			хвилин			
Wav2Lip-HD	Достатня	Нормально	Менше 10 хвилин	Високі вимоги	4,2 ГБ	Три ОС
SadTalker	Висока	Дуже зручно	Година і більше	Середні вимоги	7,6 ГБ	Три ОС
DINet	Висока	Незручно	Менше півгодини	Низькі вимоги	5.1 ГБ + 3.5 ГБ Openface	Дві ОС
MakeItTalk	Низька (генерація відео розміром 256*256)	Нормально	Менше 10 хвилин	Високі вимоги	7,2 ГБ	Дві ОС

Перетворимо в порядкові шкали (таблиця 5.2)

Таблиця 5.2 – Порядкові шкали (таблиця створена самостійно)

	Якість і ефективність	Доступність і легкість використання	Часова ефективність	Вимогливість до ресурсів ПК	Об'єм займаємої пам'яті	Багатоплатформеність
Wav2Lip	1	3	4	3	5	3
Wav2Lip-GAN	3	3	4	1	5	3
Wav2Lip-HD	3	2	4	1	4	3
SadTalker	5	5	1	3	1	3
DINet	5	2	3	5	2	2
MakeItTalk	1	3	4	1	1	2

Порівняння за Парето (таблиця 5.3):

MakeItTalk

Таблиця 5.3 – Порівняння за Парето (таблиця створена самостійно)

	Якість і ефективність	Доступність і легкість використання	Часова ефективність	Вимогливість до ресурсів ПК	Об'єм займаємої пам'яті	Багатоплатформеність
Wav2Lip	1	3	4	3	5	3
Wav2Lip - GAN	3	3	4	1	5	3
SadTalker	5	5	1	3	1	3
DINet	5	2	3	2	2	2

За Парето ми визначили 4 фаворита.

Лінійна адитивна згортка з ваговими коефіцієнтами – кожен критерій множиться на свій ваговий коефіцієнт, а потім усі зважені критерії підсумовуються і утворюють зважену цільову функцію, значення якої інтерпретується як «коефіцієнт якості» отриманого рішення. Отримана скаляризована функція максимізується в допустимому діапазоні обмежень (формула 5.2).

$$F = \sum_{j=1}^{n} \alpha_j \beta_j a_{ij}$$

де $a_j = \frac{1}{\sum_{i=1}^{m} a_{ij}}$ – нормуючі множники,

β_j – вагові коефіцієнти, сума яких дорівнює 1,

a_{ij} – значення критеріїв.

Зазначемо вагові коефіцієнти. Для цього, проведемо опитування серед експертів у цій сфері, та оберемо середні значення (таблиця 5.4).

Таблиця 5.4 – Опитування серед експертів у цій сфері (таблиця створена самостійно)

Критерій	Експерт №1	Експерт №2	Експерт №3	Середнє
Якість і ефективність	0,2	0,2	0,3	0,23
Доступність і легкість використання	0,2	0,1	0,1	0,13
Часова ефективність	0,3	0,2	0,2	0,23
Вимогливість до ресурсів ПК	0,2	0,3	0,2	0,23
Об'єм займаємої пам'яті	0,1	0,1	0,1	0,1
Багатоплатформеність	0,2	0,1	0,1	0,13

Переведемо векторний опис через вагові коефіцієнти (таблиця 5.5)

Таблиця 5.5 – Векторний опис через вагові коефіцієнти (таблиця створена самостійно)

	Якість і ефективність	Доступність і легкість використання	Часова ефективність	Вимогливість до ресурсів ПК	Об'єм займаємої пам'яті	Багатоплатформеність
Wav2Lip	0,23	0,39	0,92	0,69	0,5	0,39
Wav2Lip - GAN	0,69	0,39	0,92	0,23	0,5	0,39
SadTalker	1,15	0,65	0,23	0,69	0,1	0,39
DINet	1,15	0,26	0,69	0,46	0,2	0,26

Результати згортки представлені в таблиці 5.6.

Таблиця 5.6 – Результати згортки (таблиця створена самостійно)

Альтернативи	Сума
Wav2Lip	3,12
Wav2Lip - GAN	3,12
SadTalker	3,21

Кінець таблиці 5.6

Альтернативи	Сума
DINet	3,02

За лінійної адитивної згортки, ми визначили, що найкраща альтернатива це використання це SadTalker, бо ця модель має найвищу оцінку. За методом Парето, визначили чотирьох фаворитів.

Також визначили слабкі і сильні сторони кожного з протестованих застосунків.

5.6 Аналіз тестування різних вихідних даних

У результаті тестування застосунку з різними типами вхідних зображень були сформульовані основні рекомендації для досягнення максимальної якості відеогенерації:

- для отримання кращого результату слід обирати фотографії, зроблені в умовах м’якого, розсіяного освітлення без чітко виражених світлових плям і глибоких тіней;
- оптимальним варіантом є зображення у фронтальному ракурсі;
- необхідно використовувати фото без будь-яких об’єктів, що частково перекривають обличчя;
- важливо враховувати, що під час генерації напрям погляду визначається за орієнтацією голови, а не положенням очей на оригінальному фото;
- при роботі з неживими або стилізованими зображеннями слід зважати на можливу появу ефекту «зловісної долини», коли згенероване відео виглядає неприродно.

Крім того, у процесі дослідження були виокремлені характерні деталі, за якими можна розпізнати штучно згенероване відео або, навпаки, підтвердити автентичність медіа:

- слід звертати увагу на дрібні індивідуальні особливості шкіри, як-от родимки чи веснянки, особливо якщо відомо, що вони наявні у людини на фото;
- повна нерухомість тіла або, навпаки, надмірно активні рухи можуть свідчити про штучність відео;
- присутність яскравого та контрастного освітлення з глибокими тінями часто вказує на реальність зйомки;
- зображення, зроблені під нестандартними ракурсами (наприклад, зверху або знизу), складніше згенерувати, тому такі ракурси можуть підтверджувати справжність медіа;
- наявність об’єктів, що частково закривають обличчя, як-от рука, волосся або сторонні предмети, ускладнює генерацію та слугує додатковим критерієм справжності відео;
- зміна або спотворення геометрично правильних форм, наприклад, сережок, може вказувати на синтетичне походження відео;
- варто уважно оцінювати сталість рис обличчя та загальних пропорцій як порівняно з оригіналом, так і протягом усього відео – їхня варіативність свідчить про можливу генерацію.

ВИСНОВКИ

У цій роботі проведено дослідження методів генерації синхронізованого відеоряду на основі аудіореференсу з використанням технологій штучного інтелекту.

У ході виконання дослідження було вирішено такі завдання:

- проведено огляд предметної області та ідентифіковано основні моделі, що використовуються для синхронної відеогенерації;
- здійснено аналіз обраних моделей, визначено їхні переваги й недоліки, а також вимоги для успішного використання;
- розроблено прототипи систем для запуску і тестування застосунків на різних операційних системах;
- проведено експериментальне дослідження роботи застосунків, здійснено аналіз результатів на основі різних типів вхідних даних;
- виконано загальний аналіз отриманих результатів і сформульовано висновки.

Отримані результати можуть бути використані як основа для подальших досліджень у напрямі мультиплатформенного використання моделей генерації синхронного відеоряду. На основі проведеної роботи зроблено висновки щодо переваг і недоліків різних інструментів генерації, а також визначено вимоги до вхідних зображень для досягнення високої якості результату.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. DINet: https://github.com/natlamir/DINet (дата звернення - 10.06.2025)
2. Wav2Lip: https://github.com/Rudrabha/Wav2Lip (дата звернення - 10.06.2025)
3. MakeItTalk: https://github.com/yzhou359/MakeItTalk (дата звернення - 10.06.2025)
4. Q-Align: https://github.com/Q-Future/Q-Align (дата звернення - 10.06.2025)
5. VBench: https://github.com/Vchitect/VBench (дата звернення - 10.06.2025)
6. SyncNet: https://github.com/joonson/syncnet_python (дата звернення - 10.06.2025)
7. VQFR: https://github.com/TencentARC/VQFR (дата звернення - 10.06.2025)
8. GFPGAN: https://github.com/TencentARC/GFPGAN (дата звернення - 10.06.2025)
9. PyTorch-GAN: https://github.com/eriklindernoren/PyTorch-GAN (дата звернення - 10.06.2025)
10. SadTalker:https://github.com/OpenTalker/SadTalker (дата звернення - 10.06.2025)
11. Wav2Lip-HD:https://github.com/saifhassan/Wav2Lip-HD (дата звернення - 10.06.2025)
12. pynormalize: https://github.com/g-nie/pynormalize(дата звернення - 10.06.2025)
13. git: https://git-scm.com/ (дата звернення - 10.06.2025)
14. Python: https://www.python.org/ (дата звернення - 10.06.2025)
15. Virtualenv: https://virtualenv.pypa.io/en/latest/ (дата звернення - 10.06.2025)
16. Ffmpeg: https://ffmpeg.org/ (дата звернення - 10.06.2025)
17. Ubuntu: https://ubuntu.com/ (дата звернення - 10.06.2025)
18. Virtualbox: https://www.virtualbox.org/ (дата звернення - 10.06.2025)

19. OpenFace: https://github.com/TadasBaltrusaitis/OpenFace (дата звернення - 10.06.2025)

20. K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 484–492. https://doi.org/10.1145/3394171.3413532

21. Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, Yu Ding, 2023. DINet: Deformation Inpainting Network for Realistic Face Visually Dubbing on High Resolution Video Virtual Human Group, Netease Fuxi AI Lab, Xiamen University, Zhejiang University. https://fuxivirtualhuman.github.io/pdf/AAAI2023_FaceDubbing.pdf

22. Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, Fei Wang. 2023. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. Xi'an Jiaotong University, Tencent AI Lab, Ant Group. https://doi.org/10.1145/3394171.3413532

23. Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, Dingzeyu Li. 2020. MakeItTalk: Speaker-Aware Talking-Head Animation. SIGGRAPH Asia https://arxiv.org/abs/2004.12992

24. Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, Ziwei Liu. 2023. VBench: Comprehensive Benchmark Suite for Video Generative Models https://arxiv.org/abs/2311.17982

25. Afanasieva, I., Golian, N., Golian, V., Khovrat, A., & Onyshchenko, K. (2023). Application of Neural Networks to Identify of Fake News. COLINS (2), 346-358.

26. Каук В.І. "Generate AI - creative designer helper". // Матеріали конференції "Using new teaching methods in the publishing and printing industry ",

ХНУРЕ, 2024, с. 135-168.

27. Turuta, O.; Afanasieva, I.; Golian, N.; Golian, V.; Onyshchenko, K.; Suvorov, D.“ Audio processing methods for speech emotion recognition using machine learning". // CEUR Workshop Proceedings, 2024, http://www.scopus.com/inward/record.url?eid=2-s2.0-85197342742&partnerID=MN8TOARS

28. Saichyshyna, N.; Maksymenko, D.; Turuta, O.; Yerokhin, A.; Turuta, O.; Babii, A.“ Extension Multi30K: Multimodal Dataset for Integrated Vision and Language Research in Ukrainian". // EACL 2023 - 2nd Ukrainian Natural Language Processing Workshop, UNLP 2023 - Proceedings of the Workshop, 2023. http://www.scopus.com/inward/record.url?eid=2-s2.0-85175999944&partnerID=MN8TOARS

29. Erdem, E.; Kuyu, M.; Yagcioglu, S.; Frank, A.; Parcalabescu, L.; Babii, A.; Turuta, O.; Erdem, A.; Calixto, I.; Plank, B. et al. “ Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning". // Journal of Artificial Intelligence Research, 2022. http://www.scopus.com/inward/record.url?eid=2-s2.0-85129570397&partnerID=MN8TOARS

30. Kizitskyi, M.; Turuta, O.; Turuta, O.“ Improving Speaker Verification Model for Low-Resources Languages". // CEUR Workshop Proceedings, 2023. http://www.scopus.com/inward/record.url?eid=2-s2.0-85163101274&partnerID=MN8TOARS

31. Grynenko, A.; Turuta, O.; Kasheparov, R.; Kalynychenko, O.; Turuta, O. “ Estimation and Visualization of Webcam Eye Tracking for Text Reading". // CEUR Workshop Proceedings, 2024, http://www.scopus.com/inward/record.url?eid=2-s2.0-85210071746&partnerID=MN8TOARS

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

25. Afanasieva, I., Golian, N., Golian, V., Khovrat, A., & Onyshchenko, K. (2023). Application of Neural Networks to Identify of Fake News. COLINS (2), 346-358.

26. Каук В.І. “Generate AI - creative designer helper". // Матеріали конференції "Using new teaching methods in the publishing and printing industry ", ХНУРЕ, 2024, с. 135-168.

27. Turuta, O.; Afanasieva, I.; Golian, N.; Golian, V.; Onyshchenko, K.; Suvorov, D.“ Audio processing methods for speech emotion recognition using machine learning". // CEUR Workshop Proceedings, 2024, http://www.scopus.com/inward/record.url?eid=2-s2.0-85197342742&partnerID=MN8TOARS

28. Saichyshyna, N.; Maksymenko, D.; Turuta, O.; Yerokhin, A.; Turuta, O.; Babii, A.“ Extension Multi30K: Multimodal Dataset for Integrated Vision and Language Research in Ukrainian". // EACL 2023 - 2nd Ukrainian Natural Language Processing Workshop, UNLP 2023 - Proceedings of the Workshop, 2023. http://www.scopus.com/inward/record.url?eid=2-s2.0-85175999944&partnerID=MN8TOARS

29. Erdem, E.; Kuyu, M.; Yagcioglu, S.; Frank, A.; Parcalabescu, L.; Babii, A.; Turuta, O.; Erdem, A.; Calixto, I.; Plank, B. et al. “ Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning". // Journal of Artificial Intelligence Research, 2022. http://www.scopus.com/inward/record.url?eid=2-s2.0-85129570397&partnerID=MN8TOARS

30. Kizitskyi, M.; Turuta, O.; Turuta, O.“ Improving Speaker Verification Model for Low-Resources Languages". // CEUR Workshop Proceedings, 2023. http://www.scopus.com/inward/record.url?eid=2-s2.0-85163101274&partnerID=MN8TOARS

31. Grynenko, A.; Turuta, O.; Kasheparov, R.; Kalynychenko, O.; Turuta, O. “

Estimation and Visualization of Webcam Eye Tracking for Text Reading". // CEUR Workshop Proceedings, 2024, http://www.scopus.com/inward/record.url?eid=2-s2.0-85210071746&partnerID=MN8TOARS