https://doi.org/10.31649/1997-9266-2024-175-4-75-85

УДК 004.054:[004.032.26+004.85]

Л. Р. Кулик[1]
О. Б. Мокін[1]

МЕТОДИ ЗАБЕЗПЕЧЕННЯ КОНСИСТЕНТНОСТІ ГЕНЕРАЦІЇ В ДИФУЗІЙНИХ МОДЕЛЯХ

[1]Вінницький національний технічний університет

Досліджено проблему консистентної генерації в дифузійних моделях. Сучасні генеративні дифузійні моделі здатні створювати зображення високої точності, але підтримання консистентності між спорідненими результатами генерації залишається складним завданням. Проаналізовано ключові методи забезпечення консистентності генерації. При цьому введено додатковий тип консистентності — консистентність концепції, що дозволяє оцінити здатність моделей не тільки відтворювати існуючі стилі та об'єкти, а й генерувати абсолютно нові візуальні ідеї, з якими модель ніколи не стикалася під час навчання. Проведено аналіз наявних методів забезпечення консистентності та визначено їхні переваги та недоліки. Метод генерації на базі вхідного еталонного зображення image-to-image має перевагу в простоті реалізації. Такі методи дотренування, як DreamBooth і LoRA DreamBooth, забезпечують ширший контроль над консистентністю об'єктів. Моделі ControlNet за допомогою спеціального вхідного зображення забезпечують консистентність форми. Методи інверсії шуму, дозволяють здійснити точніший контроль та ітеративне вдосконалення підсумкових зображень за рахунок маніпуляцій з шумовим простором, що дозволяє генерувати стилістичніше та концептуально консистентні зображення. Завдяки механізму спільної уваги, що застосовується в методі StyleAligned, може забезпечуватись стилістична консистентність згенерованих зображень. Розуміння можливостей та обмежень методів забезпечення консистентності дифузійної генерації дозволяє обрати найефективніший набір інструментів відповідно до задачі. Дифузійні моделі продовжують активно розвиватися та поширюватися на нові галузі, тому досягнення надійної та універсальної консистентності в дифузійних моделях може дати шлях для креативніших та ефективніших рішень.

Ключові слова: глибоке навчання, генерація зображень, генеративні дифузійні моделі, консистентність генерації, консистентність концепції.

Вступ

Дослідження в сфері генеративного ШІ за останні роки набули захоплюючого розвитку, а роль генеративних дифузійних моделей у цьому контексті стає визначнішою [1]. На відміну від традиційних генеративних підходів, які намагаються прямо зіставити випадковий шум з вхідними даними, дифузійні моделі реалізують дієвішу стратегію. Спочатку на вхідних зображеннях застосовується прямий процес дифузії, в рамках якого відбувається поступове додавання шуму до вхідного набору даних, спотворюючи його. Після чого вже дифузійні моделі навчаються виконувати цей процес у зворотному напрямку, компенсуючи пошкоджені дані згенерованими і відновлюючи зразки вхідних даних. Ця унікальна парадигма «пряма дифузія — зворотна дифузія» має кілька переваг над традиційними методами. Вона дозволяє генерувати зображення неймовірно високої якості та точності, забезпечуючи при цьому гнучкість у керуванні рівнем деталізації під час процесу генерації. Приклади згенерованих зображень за допомогою найпродуктивнішої натепер дифузійної моделі Stable Diffusion XL (SDXL) від Stability AI [2] показано на рис. 1.

Потенційне застосування генеративних дифузійних моделей швидко розширюється. Вони знаходять своє застосування у різних творчих сферах, надаючи художникам і дизайнерам інструменти для редагування зображень, генерації текстур і матеріалів, а також створення художніх варіацій існуючих зображень [3]. У технічних сферах вони використовуються для доповнення даних — процесу, що має ключове значення для навчання інших моделей машинного навчання на ширшому спектрі даних. До того ж, проводяться дослідження застосування дифузійних моделей для малювання,

заповнення відсутніх частин зображень і навіть для створення реалістичних 3D-об'єктів [3].

Рис. 1. Приклади згенерованих зображень за допомогою дифузійної моделі SDXL

Проте широкому застосуванню дифузійних моделей заважає одна важлива проблема: забезпечення консистентності споріднених зображень під час генерування. Розглянемо приклад: модельєр хоче використати дифузійну модель для створення серії варіацій одного дизайну сукні. Щоб результат генерації дифузійної моделі був дійсно корисним, згенеровані зображення повинні бути не тільки візуально привабливими, але й підтримувати єдиний стиль і основні елементи дизайну та зображених об'єктів в усіх генераціях.

Незалежно від того, чи йдеться про дотримання єдиного художнього стилю в серії зображень, чи про створення об'єктів з точними і детальними характеристиками, які відповідають опису користувача, чи про забезпечення відповідності згенерованого зображення конкретній концепції, визначеній користувачем (навіть якщо це абсолютно нова концепція), досягнення консистентності залишається предметом постійних досліджень [4]. Це особливо важливо, оскільки в багатьох прикладних задачах згенеровані зображення мають бути не тільки візуально привабливими, але й семантично консистентними та відповідати заданій цілі.

Метою роботи є аналіз наявних методів та засобів забезпечення консистентності результатів роботи генеративних дифузійних моделей з урахуванням удосконаленого підходу до класифікації консистентності.

Дифузійні моделі: Архітектура, сильні та слабкі сторони

Дифузійні моделі працюють у два етапи: етап прямої дифузії та етап зворотної дифузії [5]. Під час прямого процесу дифузії поступово додається шум до «чистого» зразка даних. Це, по суті, спотворює оригінальне зображення, з поступовим переведенням його у стан випадкового шуму (рис. 2). Щоб формалізувати цей процес, його можна розглядати як фіксований ланцюг Маркова з кількістю кроків T, де зображення в момент часу t відображає його наступний стан в момент часу $t + 1$. Таким чином, кожен крок залежить лише від попереднього.

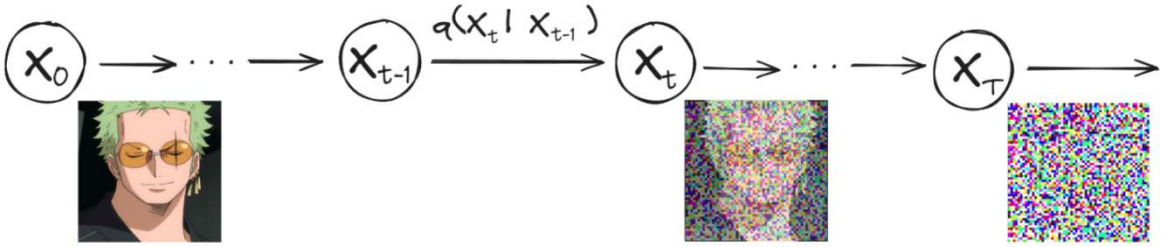

Рис. 2. Етап прямої дифузії

На етапі зворотної дифузії модель вчиться повертати цей процес назад, знешумлюючи пошкоджене зображення і відновлюючи вхідні дані (рис. 3). На відміну від процесу прямої дифузії, процес зворотної дифузії є складною обчислювальною задачею, для розв'язання якої для процесу зворотної дифузії використовуються моделі глибокого навчання, тобто штучні нейронні мережі.

Цей підхід («пряма дифузія — зворотна дифузія») має кілька переваг, однією з яких є здатність генерувати зображення високої якості та точності. Вивчаючи складні деталі та нюанси, присутні в навчальних даних, дифузійні моделі можуть створювати неймовірно реалістичні результати. В основному, цей ефект досягається за допомогою розбиття процесу зашумлення та знешумлення тренувальних даних на різні кроки, за рахунок чого відбувається природна декомпозиція тренування та відповідно спрощується задача вивчення моделлю ознак з тренувальних даних.

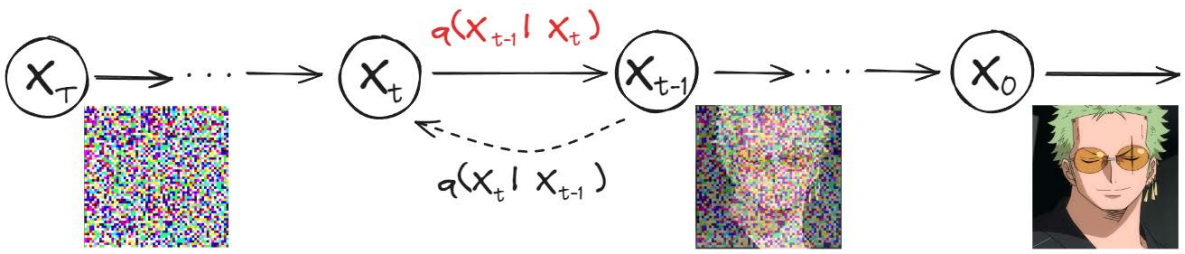

Рис. 3. Етап зворотної дифузії

До того ж, дифузійні моделі є гнучкими у керуванні рівнем деталізації під час генерації. Користувач може налаштовувати певні параметри в моделі, щоб отримати зображення від високодеталізованих і фотореалістичних до абстрактніших і стилізованіших.

Проте дифузійні моделі також мають обмеження. Основна проблема полягає в прихованому просторі моделі, тобто представленні, яке використовується для кодування вхідної інформації під час процесу зворотної дифузії. Хоча точна природа цього простору залишається на стадії вивчення [6], загальновизнано, що він не є легким в оперуванні та не піддається прямій інтерпретації. Це ускладнює безпосереднє керування конкретними аспектами згенерованого зображення, окрім як використання високорівневого текстового опису.

По суті, дифузійна модель функціонує як інструмент перетворення тексту в зображення. Вона приймає текстовий опис (підказку) як вхідні дані і використовує цю інформацію для керування процесом генерування зображень (рис. 4). Процес прямої дифузії не відбувається під час генерування. Спочатку створюється Гаусівський шум тієї ж розмірності, що має прихований простір моделі. Далі шум проходить через U-Net-мережу [7] попередньо визначену кількість кроків *T*. На кожному кроці U-Net-мережа прогнозує весь шум, присутній на зображенні. Після всіх *T* кроків формується представлення в прихованому просторі згенерованого зображення. Використовуючи декодувальну модель, яка базується на варіаційному автокодувальнику (ВАК, VAE) [8], воно переводиться з прихованого простору моделі у фінальний піксельний простір.

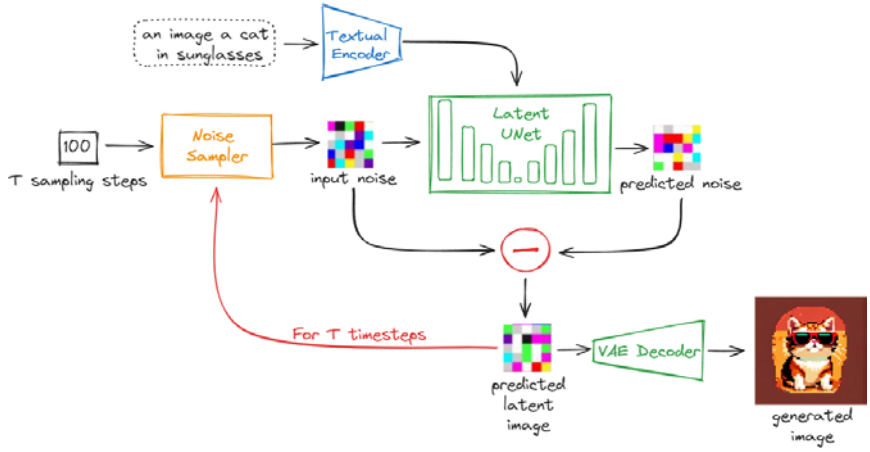

Рис. 4. Схема генерації зображення за допомогою дифузійної моделі

Методи досягнення консистентності в дифузійних моделях ґрунтуються на цій базовій функціональності. Вони або розширюють можливості самої моделі, або використовують її базові функції для конкретних цілей.

Поняття консистентності та її типи

Для розуміння та прийнятного розбиття консистентності на конкретні типи, необхідно розглянути філософське розуміння зображень та їхній зв'язок з поняттями і об'єктами. Малюнок можна розглядати як «універсальний когнітивний інструмент» для комунікації та створення візуального досвіду [9]. Зображення використовують візуальну «мову» для спілкування, покладаючись на такі елементи, як перспектива, форма та просторове розташування [10]. Хоча існують культурні та стилістичні варіації, певні аспекти цієї мови, особливо ті, що пов'язані з локальними елементами, сягають корінням у біологічний зір.

Розглядаючи поняття об’єкта в контексті створення зображень, необхідно розуміти, що це є певне поняття з набором притаманних йому ознак. Наприклад, поняття «стілець» передбачає наявність ніжок, сидіння та спинки. Проте традиційні класифікації зазвичай поділяють консистентність у створенні зображень на два основні типи: консистентність стилю та консистентність об’єкта [11]. Хоча цей підхід є відправною точкою, він не зовсім точний.

Ця двокомпонентна класифікація має кілька обмежень. По-перше, внутрішня варіативність об’єктів створює певні труднощі. «Стілець» може мати різні стилі (наприклад, обідній стілець та офісний стілець), зберігаючи при цьому свою основну функціональність. По-друге, саме поняття стилю виходить за межі суто візуальних характеристик. Воно може охоплювати мистецькі напрямки, історичні періоди або навіть стилі конкретних митців. Ці обмеження вимагають щільніших рамок для розуміння консистентності в моделях дифузії.

У світлі перелічених обмежень, концепція двоступеневої моделі візуального сприйняття [12], спонукає змінити розуміння різних типів консистентності. Ця модель припускає, що реальне 3D-сприйняття відбувається у два етапи: «бачення через погляд» для поточної фіксованої області та «інтегроване розуміння сцени» для ширшого контексту сцени (рис. 5).

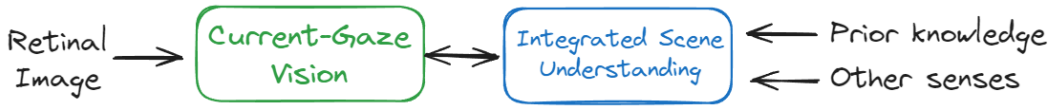

Рис. 5. Двоступенева модель візуального сприйняття

На основі цієї моделі, запропоновано три основні типи консистентності в дифузійних моделях:

1. *Консистентність стилю.* Цей тип фокусується на підтримці консистентного художнього стилю в усіх створених зображеннях. Це може включати відтворення стилю певного художника, дотримання певного мистецького напряму або кольорових палітр і мазків.

2. *Консистентність об’єкта*. Цей тип забезпечує консистентність деталей і характеристик конкретних об’єктів, зображених на згенерованих зображеннях. Основна увага приділяється збереженню основної ідентичності та характеристик об’єкта, навіть якщо він зображений під різними кутами або в різних умовах освітлення.

3. *Консистентність концепції.* Пропонується доповнити попередні два, які зазвичай беруться до уваги дослідниками, не менш важливим типом консистентності, за допомогою якого вирішується проблема створення та відтворення зображень, що відповідають визначеній користувачем концепції, навіть коли це нова концепція, якої немає в навчальних даних моделі. Консистентність концепції розширює межі дифузійних моделей, виводячи їх за рамки копіювання існуючих стилів чи об’єктів, і дозволяє генерувати абсолютно нові візуальні ідеї. Це може включати генерування зображення «велосипед з 20-ма колесами» або «хмарочос, побудований з хмарочосів», тобто концепцій, з якими модель, ймовірно, ніколи не стикалася під час навчання. Також це відноситься і до класичних концепцій, такі як «людина» або «диван», які безперечно є частиною тренувальної вибірки, але в комбінації можуть давати непередбачувані результати.

Забезпечення консистентності шляхом генерування на основі іншого зображення

Найпростішим методом досягнення консистентності об’єктів у моделях дифузії є генерація зображень за допомогою методу «image-to-image» («зображення-у-зображення») [13]. Ця техніка використовує наявне зображення як початкову точку, щоб спрямувати процес генерування до визначеного користувачем результату. Розглянемо сценарій, коли необхідно згенерувати серію зображень певного об’єкта з різних точок зору. У цьому випадку необхідно надати моделі зображення цього об’єкта як еталон. Потім модель використовує інформацію з цього еталонного зображення, щоб гарантувати, що згенеровані зображення збережуть основні характеристики.

Принцип роботи image-to-image відносно простий (рис. 6). Після отримання текстової підказки та еталонного зображення дифузійна модель перетворює інформацію із зображення у свій прихований простір. Таке перетворення, по суті, переводить інформацію про візуальні характеристики та просторові зв’язки еталонного зображення у представлення моделі, тонко впливаючи на напрямок процесу генерації зображення. Використовуючи еталонне зображення, підхід image-to-image дозволяє досягти певної консистентності об’єктів у згенерованих зображеннях.

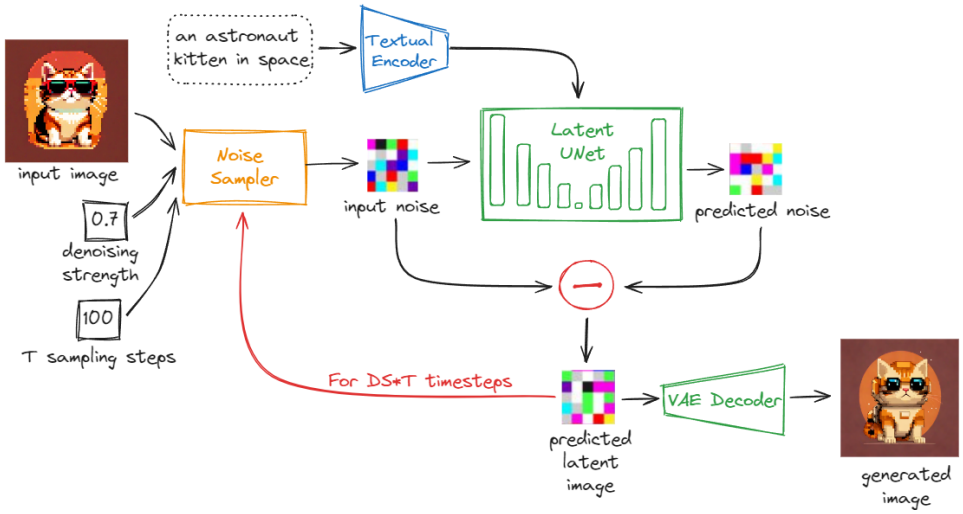

Рис. 6. Схема роботи підходу image-to-image

Проте цей підхід має низку обмежень. Однією з ключових проблем є складність досягнення точного контролю над згенерованим зображенням. Хоча еталонне зображення є орієнтиром, модель в підсумку працює в межах власних навчальних даних і внутрішніх уявлень. Це може призвести до того, що згенеровані зображення можуть дещо відхилятися від бажаного рівня консистентності. Наприклад, модель може генерувати об'єкт з дещо іншою формою рамки або іншим стилем порівняно з еталонним зображенням. Ще одне обмеження пов'язане з впливом ступеня зашумлення еталонного зображення на процес генерування. Вищий ступінь зашумлення, хоча і призводить до отримання зображень кращої якості, може також збільшити варіативність згенерованих результатів, особливо в плані дрібних деталей. Це може зменшити консистентність, особливо щодо тонких деталей, таких як подряпини або специфічні текстури, присутні на еталонному об'єкті.

Незважаючи на свої обмеження, підхід image-to-image залишається цінною основою для досягнення консистентності об'єктів. Простота і зручність використання роблять його легкодоступною технікою для різних завдань генерування зображень. Для сценаріїв, де достатнім є загальний рівень консистентності об'єктів та їхнє розташування, цей метод пропонує швидке та ефективне рішення.

Забезпечення консистентності дотренуванням дифузійних моделей

Окрім image-to-image, існують підходи, які передбачають дотренування («файнтюнинг») самої дифузійної моделі. Одним з таких підходів є DreamBooth [14], який дозволяє використати набір еталонних зображень певного об'єкта, стилю або концепції для дотренування внутрішнього представлення моделі, що дозволяє доналаштувати її для кращого генерування варіацій даних, представлених в еталонних зображеннях. Наприклад, дотренування моделі на різноманітних зображеннях певного об'єкта з різних ракурсів, умов освітлення і навіть художніх стилів. Завдяки дотренуванню за допомогою DreamBooth модель може розвинути глибше розуміння основних характеристик і потенційних варіацій, пов'язаних із зазначеним об'єктом, що дозволить їй генерувати консистентніші і точніші зображення в різних контекстах.

Проте DreamBooth також має певні обмеження. Одним з потенційних недоліків є обчислювальні витрати, пов'язані з дотренуванням всієї дифузійної моделі. Цей процес може тривати багато часу та ресурсів, особливо для складних моделей з великою кількістю параметрів. Більше того, DreamBooth насамперед фокусується на консистентності об'єктів і може не дуже добре підходити для досягнення консистентності складних концепцій. Процес дотренування, по суті, підсилює наявні знання в моделі на основі еталонних зображень.

Деякі з цих обмежень дозволяє усунути LoRA (Low-Rank Adaptation) DreamBooth [15], тобто варіація DreamBooth, яка використовує техніку адаптації базової моделі з використанням моделі нижчого рангу. По суті, це означає, що на попередньо навчену дифузійну модель накладається менша модель. Ця додаткова модель фокусується на врахуванні специфічних варіацій і характеристик, пов'язаних з тренувальними даними. При цьому сама основна дифузійна модель залишається практично незмінною.

Підхід LoRA DreamBooth має кілька переваг. По-перше, він є ефективнішим в обчислювальному плані порівняно з точним налаштуванням всієї моделі дифузії. Це робить його практичнішим

варіантом для ситуацій, коли ресурси обмежені. По-друге, він дозволяє потенційно швидше адаптуватися до нових об'єктів або простих концепцій, оскільки оновлювати потрібно лише відносно невелику LoRA модель. Однак ця ефективність несе за собою і відповідні недоліки. Через свою низько-рангову природу LoRA DreamBooth має деякі обмеження щодо відтворення консистентної концепції. Хоча цей підхід може досягти хороших результатів для завдань консистентності об'єктів, його здатність працювати з абсолютно новими та складними концепціями залишається обмеженою. Наприклад, модель адаптації однаково може не охопити всю складність нової концепції, якщо вона передбачає поєднання елементів, яких немає в еталонному наборі.

Ось як DreamBooth працює з практичного погляду:

1. Підготовка даних: збір еталонних зображень, що демонструють об'єкт, стиль або концепцію.

2. Дотренування: дифузійна модель проходить процес дотренування з використанням цих еталонних зображень. В випадку тренування LoRA DreamBooth — це не змінює всю модель, а додає лише меншу модель (модуль), спеціально сфокусовану на наданому об'єкті/стилі/концепції.

3. Кодування сутності: під час дотренування модель аналізує еталонні зображення, виділяючи ключові особливості та варіації, пов'язані з об'єктом/концептом. По суті, вона вивчає ознаки зображень.

4. Генерація з підказками: в процесі тренування кожне еталонне зображення має текстову підказку, та, зазвичай, закодований ідентифікатор, що застосовується для асоціативного тренування, тобто асоціація ідентифікатора з еталонними прикладами. Модель використовує ці підказки та кодовану сутність зображень, щоб керувати процесом генерації зображень, наближаючи генерацію до еталонних тренувальних прикладів.

Забезпечення консистентності форми об'єкта

Розглянуті вище методи пропонують цінні інструменти для досягнення консистентності об'єктів та стилів. Однак досягнення консистентності часто виходить за рамки лише відтворення коректних деталей об'єкта/стилю і поширюється на контроль загальної форми об'єкта на згенерованому зображенні. І тут варто згадати ще один потужний інструмент — модель ControlNet [16].

Модель ControlNet працює шляхом включення спеціального вхідного зображення разом з текстовою підказкою у процес зворотної дифузії. Це вхідне зображення діє як «спрямовувальна форма», впливаючи на загальну форму/контури/силует об'єкта, що генерується. Розглянемо на прикладі. Нехай є необхідність згенерувати серію зображень кота. У той час як DreamBooth може забезпечити консистентність котячих рис (вуса, вуха тощо) об'єкта, який генерується, ControlNet дозволяє контролювати загальну форму та представлення кота на всіх згенерованих зображеннях. Таким чином можна використати контрольне зображення кота, щоб згенерувати кілька таких же зображень, тільки, наприклад, в іншій стилістиці.

Ключова перевага ControlNet полягає в можливості маніпулювати формою об'єктів різними способами. Це дозволяє контролювати позу, поставу та загальний силует об'єкта на згенерованому зображенні, а також забезпечує гнучкість у досягненні однакових форм на серії зображень, навіть якщо йдеться про складні об'єкти або абстрактні поняття.

Проте важливо зауважити, що ControlNet сам по собі не є повноцінним рішенням для консистентності об'єктів. Хоча він добре справляється з маніпулюванням формою, він не контролює безпосередньо властиві об'єкта та його риси. Цей підхід є по суті інструментом, зосередженим на «як» (форма), а не на «що» (властивості) об'єкта. Наприклад, використання еталонного зображення кота не обов'язково забезпечить консистентність деталей, таких як текстура хутра або колір очей. Тому ControlNet слід розглядати як допоміжний інструмент для консистентної генерації. Він доповнює наявні методи, дозволяючи контролювати форму об'єктів або навіть стильового чи концептуального представлення. Це дозволяє генерувати варіації в межах певної форми, потенційно змінюючи стиль і відображення об'єкта на згенерованому зображенні.

ControlNet використовує попередньо навчений модуль під конкретну дифузійну модель, спеціально розроблену для управління формою. Під час процесу генерації в загальну модель подається текстова підказка і еталонне зображення на вхід. Дифузійна модель обробляє текстову підказку, щоб зрозуміти бажаний зміст, тоді як керуюча мережа аналізує еталонне зображення, щоб виділити інформацію про форму. Потім ця інформація про форму вводиться в процес зворотної дифузії, впливаючи на те, як модель генерує форму об'єкта на зображенні.

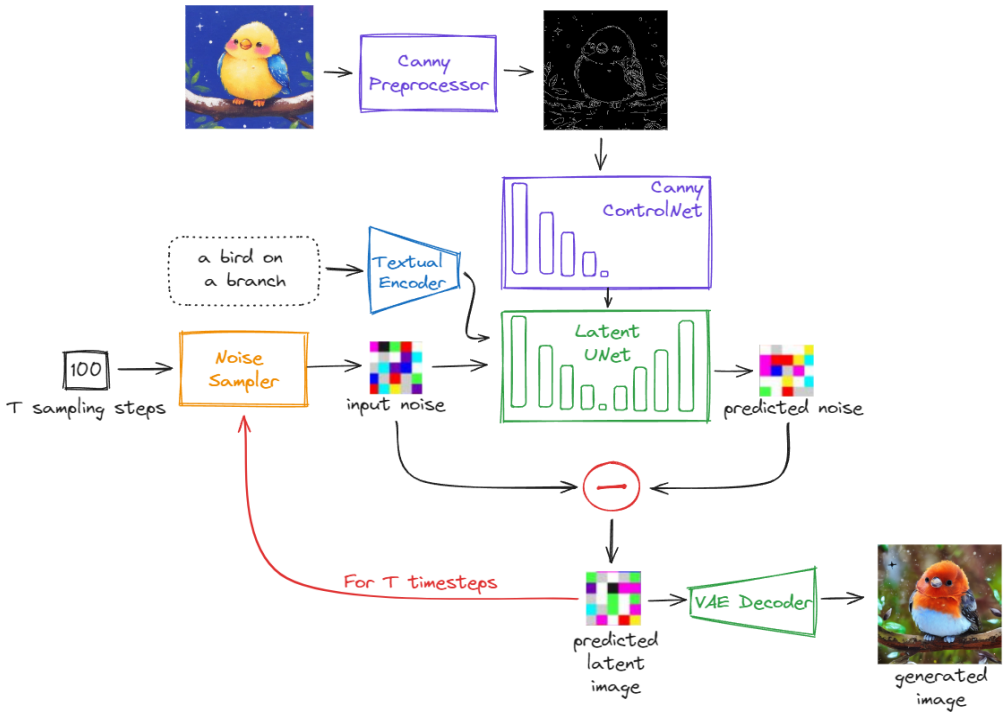

Рис. 7. Схема роботи моделі ControlNet для границь/контурів в процесі зворотної дифузії

Користувачу доступно кілька попередньо навчених моделей ControlNet, кожна з яких зосереджена на певному типі керування формою [16]. Наприклад, деякі моделі чудово справляються з керуванням границь/контурів об'єктів, тоді як інші спеціалізуються на маніпуляціях з людськими позами або інформацією про глибину. Приклад роботи дифузійної моделі з ControlNet для відтворення зображень з урахуванням заздалегідь визначених границь/контурів показанио на рис. 7. Така різноманітність дозволяє вибрати відповідну модель ControlNet, враховуючи конкретні потреби полішення консистентності форми.

Забезпечення консистентності за допомогою інверсії шуму

Хоча підхід image-to-image пропонує базовий рівень консистентності завдяки використанню еталонного зображення, він в основному ґрунтується на реальному шумі, який додається в процесі зворотної дифузії. Однак для досягнення надійнішої та контрольованішої консистентності існують методи інверсії шуму [17]. Для застосування цих методів використовують еталонне зображення і перетворюють його на специфічне «шумове» представлення в прихованому просторі вибраної дифузійної моделі. Це представлення, коли воно бере участь у процесі зворотної дифузії, дозволяє повною мірою реконструювати оригінальне зображення.

Методи інверсії шуму надають широкі можливості для досягнення консистентності в моделях дифузії, наприклад:

1. Цілеспрямовані модифікації: маніпулюючи представленням шуму, можна вибірково змінювати певні аспекти зображення, зберігаючи загальний стиль і структуру. Це дозволяє цілеспрямовано контролювати консистентність, даючи змогу вносити зміни і зберігаючи основні візуальні елементи.

2. Кращий контроль: порівняно з image-to-image інверсія шуму пропонує тонший контроль над процесом консистентності. Є можливість глибше зануритися в представлення зображення в рамках прихованого простору моделі, потенційно впливаючи на певні особливості або об'єкти на зображенні в точніший спосіб.

3. Ітеративне вдосконалення: інверсія шуму дозволяє застосувати ітеративний підхід до консистентності, тобто в процесі виконання інверсії шуму маніпулювати представленням шуму, а потім повернути його назад у процес дифузії для уточнення згенерованого зображення. Цей ітеративний цикл дозволяє поступово досягти бажаного рівня консистентності.

Один з популярних методів прикладного застосування інверсії шуму для розв'язання задачі консистентності, описаний у статті «An Edit Friendly DDPM Noise Space: Inversion and Manipulations» («Редаговані інтуїтивні шумові простори DDPM: інверсія та маніпуляції») [18]. Цей

метод фокусується на маніпулюванні шумовим простором DDPM (Denoising Diffusion Probabilistic Models, «знешумлювальні ймовірнісні дифузійні моделі») для досягнення консистентності за рахунок маніпуляцій з текстовими підказками в процесі генерування.

Основна перевага цього методу полягає в його здатності створювати зручний для редагування простір шуму. Цей простір дозволяє легко маніпулювати окремими деталями або об'єктами зображення, зберігаючи при цьому загальний стиль. Уявіть, що у вас є зім'ятий малюнок, на якому ви можете легко розгладити певні зморшки (небажані деталі), не впливаючи на решту зображення.

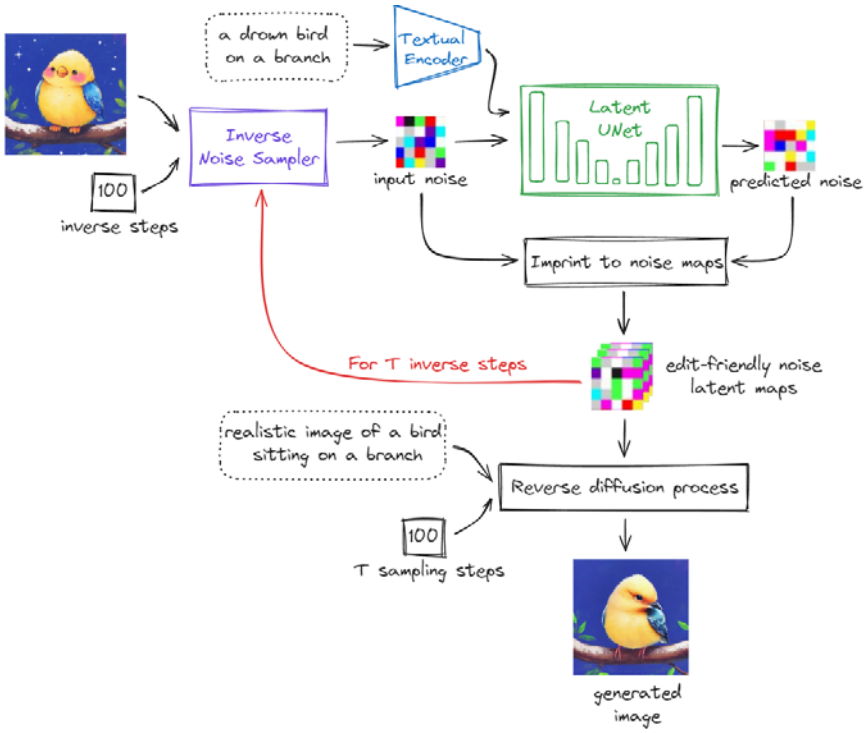

Рис. 8. Схема роботи вибраного для прикладу методу інверсії шуму

Ось як працює цей метод (рис. 8):

1. Інверсія: метод отримує вхідне зображення та інвертує його в простір шуму DDPM. Це створює шумове представлення, яке кодує інформацію зображення.

2. Редагування: користувач може впливати на це шумове представлення для внесення певних змін за рахунок маніпуляцій з текстовими підказками. Це може включати додавання нових об'єктів, модифікацію наявних деталей або навіть видалення небажаних елементів.

3. Розповсюдження та вдосконалення: відредаговане шумове представлення повертається до процесу зворотної дифузії, тобто безпосередньої генерації зображення з використанням дифузійної моделі, генеруючи зображення, яке відображає внесені зміни, зберігаючи при цьому стиль оригінального зображення.

Простота і зручність цього методу роблять його цінним інструментом для досягнення консистентності результатів роботи генеративних дифузійних моделей. Він дозволяє контролювати певні аспекти зображення, залишаючись в рамках оригінального стилю.

Забезпечення консистентності стилю за допомогою механізму спільної уваги

Хоча методи дотренування та інверсії шуму певною мірою вирішують проблему консистентності об'єктів і концепцій, вони можуть мати проблеми з підтриманням єдиного стилю в серії згенерованих зображень. Метод StyleAligned [19] вирішує цю проблему, зосереджуючись на забезпеченні стилістичної консистентності створених зображень за допомогою підходу, який отримав назву Shared Attention, тобто «спільна увага».

StyleAligned працює в рамках процесу зворотної дифузії. Нагадаємо, що під час зворотної дифузії шум поступово видаляється з прихованого шумового представлення, щоб сформувати зображення. StyleAligned вводить у цей процес етап «спільної уваги», принцип дії якого розглянемо на прикладі. Нехай є група художників (дифузійних моделей), що намагається писати картини у схожому стилі. У StyleAligned один художник (еталонне зображення) виступає в ролі наставника,

а інші (цільові зображення) звертають увагу на процес створення та прийоми наставника (інформацію про стиль) під час власного процесу написання картини (створення зображення).

Ця спільна увага досягається за допомогою спеціального механізму, який називається AdaIN [20], за допомогою якого цільові зображення «дивляться» на еталонне зображення, щоб отримати інформацію про стиль. Потім ця інформація використовується для керування процесом генерування цільових зображень, забезпечуючи дотримання того ж стилю, що й у еталонному зображенні. Приклади генерацій з використанням цього методу подано на рис. 9.

Рис. 9. Приклади зображень, згенерованих за допомогою методу StyleAligned

З переваг методу StyleAligned можна виділити:

1. Консистентність стилю: StyleAligned забезпечує консистентність стилю без необхідності складних налаштувань або ручного втручання. Це робить його зручним та ефективним підходом.

2. Високоякісна та різноманітна генерація: метод створює високоякісні зображення, які відповідають еталонному стилю, навіть в рамках різних текстових підказок. Це дає змогу створити серію зображень, які зберігають єдиний стилістичний вигляд за умови змінного змісту.

З недоліків виділяється лише пряма орієнтація на концепцію та стиль. StyleAligned чудово підтримує стилістичну консистентність, але він насамперед підходить для копіювання концептуальної або стилістичної теми з еталонного зображення. Підхід не має прямого відношення до консистентності об’єктів. Іншими словами, його неможливо використати для прямого копіювання об’єктів з еталонного зображення у згенеровані зображення. Це обмеження випливає з природи процесу розподіленої уваги. Цільові зображення фокусуються на загальній стильовій інформації, а не на конкретних об’єктах, показаних на еталонному зображенні.

Висновки

Дифузійні моделі зробили революцію у напрямку генерування зображень, даючи можливість створювати високоточні зображення з текстових описів. Проте забезпечення консистентності між серіями згенерованих зображень залишається важливою та багатошаровою проблемою.

Для ретельнішого та якіснішого аналізу введено поняття консистентності концепції, яке дозволяє оцінити здатність дифузійної моделі генерувати абсолютно нові візуальні ідеї, з якими модель ніколи не стикалася під час навчання.

Проаналізовано найактуальніші методи забезпечення консистентності результатів роботи дифузійних моделей. Спочатку розглянуто підхід image-to-image з використанням еталонних зображень. Далі — методи дотренування DreamBooth і LoRA DreamBooth, які за властивостями мають глибшу консистентність для об’єктів, але з обмеженнями в роботі з абсолютно новими концепціями. Моделі ControlNet є потужними інструментами для керування консистентністю форми, а метод інверсії шуму дозволяє здійснити точніший контроль та ітеративне доопрацювання генерації. Нарешті, в методі StyleAligned задіяно механізм спільної уваги для забезпечення стилістичної консистентності згенерованих зображень. Хоча кожен метод має свої сильні та слабкі сторони, разом вони прокладають шлях до досягнення різноманітних форм консистентності результатів роботи генеративних дифузійних моделей.

Виконаний огляд та запропоновані висновки дають цінну інформацію для дослідників і практиків, які працюють з моделями дифузії. Розуміючи можливості та обмеження існуючих методів забезпечення консистентності, розробники можуть робити обґрунтований вибір під час вибору найвідповіднішого методу для своїх конкретних потреб. Майбутні дослідження можуть бути направлені на вдосконалення методів дотренування для ефективнішої обробки складних концепцій, а також розробку нових методів консистентності, які охоплюють ширший спектр консистентності,

ніж просте відтворення об’єктів або стилів. Зрештою, досягнення надійної та універсальної консистентності в дифузійних моделях може дати шлях для креативніших та ефективніших застосувань у різних галузях.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Chenshuang Zhang, Chaoning Zhang, et al., “Text-to-image Diffusion Models in Generative AI: A Survey,” in *arXiv e-prints*, 2023. [Online]. Available: https://arxiv.org/abs/2303.07909 . Accessed on: April 29, 2024.

[2] Dustin Podell, Zion English, et al., “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis,” in *arXiv e-prints*, 2023. [Online]. Available: https://arxiv.org/abs/2307.01952 . Accessed on: April 29, 2024.

[3] Ling Yang, Zhilong Zhang, et al., “Diffusion Models: A Comprehensive Survey of Methods and Applications,” in *arXiv e-prints*, 2022. [Online]. Available: https://arxiv.org/abs/2209.00796 . Accessed on: April 29, 2024.

[4] Omri Avrahami, Amir Hertz, et al., “The Chosen One: Consistent Characters in Text-to-Image Diffusion Models,” in *arXiv e-prints*, 2023. [Online]. Available: https://arxiv.org/abs/2311.10093 . Accessed on: April 29, 2024.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel, “Denoising Diffusion Probabilistic Models,” in *arXiv e-prints*, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239 . Accessed on: April 29, 2024.

[6] Yong-Hyun Park, Mingi Kwon, et al., “Understanding the Latent Space of Diffusion Models through the Lens of Riemannian Geometry,” in *arXiv e-prints*, 2023. [Online]. Available: https://arxiv.org/abs/2307.12868 . Accessed on: April 29, 2024.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *arXiv e-prints*, 2015. [Online]. Available: https://arxiv.org/abs/1505.04597 . Accessed on: April 29, 2024.

[8] Diederik P. Kingma, Max Welling, et al., “An Introduction to Variational Autoencoders,” in *arXiv e-prints*, 2019. [Online]. Available: https://arxiv.org/abs/1906.02691 . Accessed on: April 29, 2024.

[9] Fan Judith. E., Bainbridge Wilma. A., et al, “Drawing as a versatile cognitive tool,” *Nature Reviews Psychology*, 2023. https://doi.org/10.1038/s44159-023-00212-w .

[10] G. Greenberg, “Semantics of pictorial space,” *Springer Link*, 2021. https://doi.org/10.1007/s13164-020-00513-6 .

[11] Gihyun Kwon, and Jong Chul Ye, “Diffusion-based Image Translation using Disentangled Style and Content Representation,” in *arXiv e-prints*, 2022. [Online]. Available: https://arxiv.org/abs/2209.15264 . Accessed on: April 29, 2024.

[12] Aaron Hertzmann, “Toward a theory of perspective perception in pictures,” *Journal of Vision*, 2024. https://doi.org/10.1167/jov.24.4.23 .

[13] Chenlin Meng, Yutong He, et al., “SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations,” in *arXiv e-prints*, 2021. [Online]. Available: https://arxiv.org/abs/2108.01073 . Accessed on: April 29, 2024.

[14] Nataniel Ruiz, Yuanzhen Li, et al., “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation,” in *arXiv e-prints*, 2022. [Online]. Available: https://arxiv.org/abs/2208.12242. Accessed on: April 29, 2024.

[15] Edward J. Hu, Yelong Shen, et al., “LoRA: Low-Rank Adaptation of Large Language Models,” in *arXiv e-prints*, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685 . Accessed on: April 29, 2024.

[16] Lvmin Zhang, Anyi Rao, et al., “Adding Conditional Control to Text-to-Image Diffusion Models,” in *arXiv e-prints*, 2023. [Online]. Available: https://arxiv.org/abs/2302.05543. Accessed on: April 29, 2024.

[17] Ron Mokady, Amir Hertz, et al., “Null-text Inversion for Editing Real Images using Guided Diffusion Models,” in *arXiv e-prints*, 2022. [Online]. Available: https://arxiv.org/abs/2211.09794 . Accessed on: April 29, 2024.

[18] Inbar Huberman-Spiegelglas, Vladimir Kulikov, et al., “An Edit Friendly DDPM Noise Space: Inversion and Manipulations,” in *arXiv e-prints*, 2023. [Online]. Available: https://arxiv.org/abs/2304.06140 . Accessed on: April 29, 2024.

[19] Amir Hertz, Andrey Voynov, et al., “Style Aligned Image Generation via Shared Attention,” in *arXiv e-prints*, 2023. [Online]. Available: https://arxiv.org/abs/2312.02133. Accessed on: April 29, 2024.

[20] Xun Huang, Serge Belongie, “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization,” in *arXiv e-prints*, 2017. [Online]. Available: https://arxiv.org/abs/1703.06868 . Accessed on: April 29, 2024.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 29.05.2024

Кулик Леонід Русланович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: leonidkulik2707@gmail.com ;

Мокін Олександр Борисович — д-р техн. наук, професор, професор кафедри системного аналізу та інформаційних технологій, e-mail: abmokin@gmail.com .

Вінницький національний технічний університет, Вінниця

L. R. Kulyk[1]
O. B. Mokin[1]

Methods for Ensuring Consistent Generation in Diffusion Models

[1]Vinnytsia National Technical University

The article investigates the problem of consistent generation in diffusion models. Modern generative diffusion models are capable of creating high-precision images, but maintaining the consistency between the related generation results remains a challenging task. The key methods for ensuring generation consistency are analyzed. Additionally, a new type of consistency is introduced — conceptual consistency, which allows for assessing the models' ability not only to reproduce existing styles and objects but also to generate entirely new visual ideas that the model has never encountered during training. The existing methods for ensuring consistency are analyzed, and their advantages and disadvantages are identified. The image-to-image generation method based on an input reference image has the advantage of simplicity in implementation. Fine-tuning methods like DreamBooth and LoRA DreamBooth provide broader control over object consistency. ControlNet models ensure shape consistency using a special input image that serves as a guide shape in the reverse diffusion process. Noise inversion methods allow for more precise control and iterative refinement of the resulting images through manipulations with the noise space, enabling the generation of more stylistically and conceptually consistent images. The StyleAligned method, using a shared attention mechanism, can ensure the stylistic consistency of generated images. Understanding the capabilities and limitations of methods for ensuring diffusion generation consistency allows for selecting the most effective set of tools according to the task at hand. Diffusion models continue to evolve and expand into new areas, so achieving reliable and universal consistency in diffusion models could pave the way for even more creative and effective solutions.

Keywords: deep learning, image generation, generative diffusion models, generation consistency, conceptual consistency.

Kulyk Leonid R. — Post-Graduate Student with the Chair of System Analysis and Information Technologies, e-mail: leonidkulik2707@gmail.com ;
Mokin Oleksandr B. — Dr. Sc. (Eng.), Professor, Professor with the Chair of System Analysis and Information Technologies, e-mail: abmokin@gmail.com