УДК 004.89:004.932.72:004.738.5
DOI: 10.26102/2310-6018/2026.53.2.005

Практические аспекты построения приватных мультимодальных генеративных моделей: методы, ограничения и инструменты

Е.В. Ледовская✉

МИРЭА – Российский технологический университет, Москва, Российская Федерация

Резюме. В статье рассматривается актуальная проблема разработки генеративных систем искусственного интеллекта, способных работать с разнородными данными (текст, изображение, аудио) без нарушения конфиденциальности исходных обучающих наборов. Цель работы – систематизировать и представить с практической точки зрения современные методы обеспечения приватности, применимые к мультимодальным архитектурам. Основное внимание уделяется технологиям дифференциальной приватности и федеративного обучения, их адаптации и композиции для работы со сложными данными. В статье анализируются фундаментальные компромиссы между качеством генерации, вычислительной сложностью и уровнем гарантий конфиденциальности, с которыми сталкивается разработчик на практике. Приводятся примеры существующих программных фреймворков и даются рекомендации по выбору стратегии защиты в зависимости от типа решаемой задачи и характера мультимодальных данных. Дополнительно обсуждаются практические аспекты интеграции приватных механизмов в тренировочные циклы, оценка накопленного бюджета конфиденциальности, а также потенциальные направления развития инструментов для повышения эффективности и надежности AI-систем. Отдельное внимание уделяется вопросам согласования модальностей и оптимизации компромисса между уровнем приватности и качеством генерации. Представленные рекомендации и примеры реализации могут служить руководством для инженеров и исследователей при разработке реальных мультимодальных систем, соответствующих современным требованиям безопасности и этики. Материал статьи ориентирован на исследователей и инженеров в области машинного обучения, занимающихся созданием отвечающих этическим и регуляторным требованиям AI-систем.

Ключевые слова: генеративные модели, мультимодальное машинное обучение, конфиденциальность данных, дифференциальная приватность (DP), федеративное обучение (FL), компромисс приватность-качество, фреймворки машинного обучения, устойчивые AI-системы.

Для цитирования: Ледовская Е.В. Практические аспекты построения приватных мультимодальных генеративных моделей: методы, ограничения и инструменты. *Моделирование, оптимизация и информационные технологии.* 2026;14(2). URL: https://moitvivt.ru/ru/journal/article?id=2169 DOI: 10.26102/2310-6018/2026.53.2.005

Practical aspects of building private multimodal generative models: methods, constraints, and tools

E.V. Ledovskaya✉

MIREA – Russian Technological University, Moscow, the Russian Federation

Abstract. The article addresses the pressing issue of developing generative artificial intelligence systems capable of working with heterogeneous data (text, images, audio) without compromising the privacy of the underlying training datasets. The aim of the study is to systematize and present, from a practical perspective, current methods for ensuring privacy applicable to multimodal architectures. Particular attention is paid to differential privacy and federated learning technologies, their adaptation, and their

combination for working with complex data. The article analyzes fundamental trade-offs between generation quality, computational complexity, and the level of privacy guarantees faced by developers in practice. Examples of existing software frameworks are provided, along with recommendations for selecting protection strategies depending on the type of task and the nature of the multimodal data. Practical aspects of integrating privacy mechanisms into training cycles, assessing the accumulated privacy budget, and potential directions for developing tools to enhance the efficiency and reliability of AI systems are additionally discussed. Special attention is given to issues of modality alignment and optimizing the trade-off between privacy level and generation quality. The presented recommendations and implementation examples can serve as a guide for engineers and researchers in developing real-world multimodal systems that meet contemporary security and ethical requirements. The material of the article is intended for researchers and engineers in the field of machine learning who are engaged in creating AI systems that comply with ethical and regulatory standards.

Keywords: generative models, multimodal machine learning, data privacy, differential privacy (DP), federated learning (FL), privacy-utility trade-off, machine learning frameworks, trustworthy AI systems.

For citation: Ledovskaya E.V. Practical aspects of building private multimodal generative models: methods, constraints, and tools. *Modeling, Optimization and Information Technology*. 2026;14(2). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2169 DOI: 10.26102/2310-6018/2026.53.2.005

Введение

Бурное развитие генеративных моделей машинного обучения привело к их широкому внедрению в прикладные и коммерческие системы. Архитектуры класса GPT, диффузионные модели изображений (Stable Diffusion, DALL·E) и мультимодальные модели нового поколения (GPT-4, Sora, Gemini) эффективно работают с текстом, изображениями, аудио и видео. Применение генеративных моделей в таких областях, как анализ пользовательского контента, автоматизация бизнес-процессов, медицина и образование, связано с необходимостью обработки конфиденциальных данных.

Параллельно ужесточается нормативное регулирование защиты данных: GDPR и подобные инициативы предполагают обеспечение безопасного сбора, хранения и обработки персональных данных, исключающее их несанкционированное раскрытие, а также предотвращающее восстановление данных на основе выходных результатов алгоритмов. Для генеративных моделей это означает необходимость предотвращения утечек обучающих данных и меморизации чувствительной информации.

Особую сложность представляют мультимодальные системы, где гетерогенные данные формируют сложные латентные взаимосвязи. Это создает риски косвенной утечки: информация из одной модальности может быть восстановлена через другую, а стандартные методы анонимизации оказываются недостаточными [1].

Практическую значимость приобретают методы приватности, адаптированные к мультимодальному обучению, прежде всего дифференциальная приватность и федеративное обучение, ограничивающие утечки на уровне параметров и процесса обучения. Однако их применение сопровождается ростом вычислительных затрат, снижением качества генерации и усложнением архитектуры.

Цель статьи – представить практико-ориентированный обзор методов и инструментов для построения приватных мультимодальных генеративных моделей, анализировать ограничения и компромиссы, а также показать примеры существующих программных решений в инженерных проектах.

Краткий перечень ключевых задач, решаемых в работе:

– определение системы. Четкое инженерное описание мультимодальной генеративной модели как системы для обработки и генерации согласованных данных разных типов на основе единого латентного пространства;

– анализ угроз. Выявление специфических рисков утечки данных для генеративных моделей, включая меморизацию, кросс-модальные зависимости и условную генерацию;

– формулировка требований. Определение практических требований к приватности: предотвращение определения участия записи в обучении, связывания выходов с источниками и восстановления исходных данных;

– инженерная постановка. Установка трех условий реализации: сохранение полезности модели (качества генерации), обеспечение формальных гарантий приватности, поддержание приемлемых вычислительных затрат;

– исследование методов. Анализ и адаптация основных методов обеспечения приватности (дифференциальная приватность, федеративное обучение, их гибрид) для специфики мультимодальных генеративных моделей;

– сравнительный анализ. Сопоставление достоинств, недостатков и компромиссов рассмотренных методов с точки зрения гарантий приватности, качества модели и ресурсных затрат.

Материалы и методы

В работе использованы результаты исследований Naseri et al. [2], Sun et al. [3], Feretzakis et al. [1] и Rafi et al. [4], касающиеся реализации дифференциальной приватности (DP) и федеративного обучения (FL) [5] в мультимодальных генеративных моделях. Для экспериментов рассматривались архитектуры с текстовыми, визуальными и аудиоданными, интегрированные с механизмами DP-SGD и распределенными схемами FL с учетом накопленного бюджета приватности. Эффективность методов оценивалась по компромиссу между качеством генерации, вычислительными затратами и формальными гарантиями конфиденциальности с использованием библиотек Opacus, TensorFlow Privacy и фреймворков FL (NVIDIA FLARE, Flower, PySyft).

Результаты

Постановка практической задачи. В данной работе мультимодальная генеративная модель рассматривается как инженерная система, предназначенная для совместной обработки и генерации данных разных модальностей на основе единого или согласованного латентного представления. На практике наиболее распространены архитектуры, генерирующие изображения по тексту или синтезирующие аудиовизуальный контент, с использованием специализированных энкодеров для каждой модальности и механизмов их согласования в процессе обучения. Обучение таких моделей требует доступа к крупным мультимодальным корпусам с реальными пользовательскими данными, что делает обеспечение конфиденциальности неотъемлемой частью проектирования.

Защита данных в генеративном моделировании означает предотвращение извлечения информации о конкретных обучающих примерах из параметров модели, промежуточных представлений или сгенерированных выходов. В отличие от классических задач анализа, угрозы здесь проявляются не только через прямой доступ к данным, но и через косвенные каналы, связанные с меморизацией, латентными зависимостями между модальностями и поведением модели при условной генерации. Для мультимодальных систем это увеличивает риск восстановления сведений одной модальности через другую, что осложняет применение стандартной анонимизации [4].

Практические требования к приватности включают ограничение возможности внешнего наблюдателя определять участие отдельных записей в обучении, устанавливать связь между результатами генерации и источниками данных, а также восстанавливать элементы исходных примеров. Эти требования действуют как для

централизованного, так и для распределенного обучения на стороне клиента. Конфиденциальность рассматривается как свойство всей обучающей процедуры и итоговой модели.

С инженерной точки зрения успешная реализация задачи достигается при соблюдении трех условий: сохранении достаточной полезности модели (качество и согласованность сгенерированного контента), обеспечении формальных гарантий приватности (например, через параметры дифференциальной приватности) и поддержании приемлемых вычислительных и ресурсных затрат. Усиление приватности неизбежно увеличивает вычислительную нагрузку и усложняет оптимизацию, особенно для мультимодальных моделей с большим числом параметров и сложными схемами выравнивания представлений.

Арсенал практических методов обеспечения приватности. Дифференциальная приватность в контексте обучения генеративных моделей представляет собой формальный механизм ограничения влияния отдельных обучающих записей на результаты алгоритма посредством контролируемого внесения случайного шума. В практическом применении этот подход описывается через параметры ε и δ: ε задает верхнюю границу различимости распределений выходов при наличии или отсутствии одной записи в обучающем множестве, а δ отражает маловероятные отклонения от строгой ε-ограниченности. Здесь параметр δ – это незначительная вероятность нарушения формальной гарантии ε-приватности. Инженерная реализация DP в системах машинного обучения достигается путем модификации обучающего процесса или представлений данных; ключевой целью является обеспечение измеримой и учтенной деградации утечки информации при допустимом снижении полезности модели [2].

На уровне оптимизации наиболее распространенной практикой является дифференциально-приватный стохастический градиентный спуск. В DP-SGD вычисляемые на мини-пакетах градиенты сначала нормируются (clipping) для ограничения влияния отдельных примеров, после чего добавляется шум, распределенный согласно выбранной механизму (обычно гауссовскому), и только затем происходит обновление параметров. Такой механизм управляет вкладом каждой записи в обучение и обеспечивает учет приватности через механизмы подсчета (privacy accounting) по итерациям. Альтернативный подход заключается во внесении шума на уровне выходов энкодера или в латентных представлениях; в этом случае приватность контролируется до этапа агрегации или декодирования, что дает иную точку компромисса между потерями полезности и степенью защиты [3].

Специфика мультимодальных архитектур обусловливает дополнительные архитектурные и прикладные решения по месту и способу внесения шума. В практических системах различают стратегию независимой обработки модальностей, когда каждый энкодер снабжается собственным механизмом приватности, и стратегию внесения шума в общее латентное пространство, формируемое после согласования модальностей. Независимая обработка упрощает таргетированную настройку уровня шума под особенности конкретной модальности и может снизить разрушительное влияние на внутреннее качество представлений, тогда как шум, примененный в общем латентном пространстве, обеспечивает защиту кросс-модальных связей, но более прямо сказывается на способности модели корректно связывать и генерировать контент между модальностями. Выбор подхода определяется архитектурой, доступностью вычислительных ресурсов и требуемым уровнем качества кросс-модальной генерации.

Федеративное обучение представляет собой иной класс мер, ориентированных на архитектуру развертывания: обучение организуется распределенно, исходные данные хранятся локально на устройствах или в локальных организациях, а обмен осуществляется через параметры, градиенты или иные агрегируемые объемы информации. Принцип «данные не покидают устройство» означает, что центральный

сервер получает только агрегированные обновления, а не исходные примеры. На практике одной из наиболее распространенных схем агрегации является FedAvg, при которой сервер усредняет локальные обновления, поступающие от множества клиентов, для формирования глобальной модели. Такая схема снижает риск прямого доступа к обучающим данным, однако не обеспечивает формальных приватных гарантий на уровне отдельных записей: локальные обновления и градиенты содержат статистическую информацию, которая при неблагоприятных условиях (например, малое количество клиентов, сильная неоднородность данных) может быть использована для реконструкции или вывода о локальных примерах.

Комбинация федеративного обучения и дифференциальной приватности представляет собой практический тренд в инженерной реализации приватных мультимодальных систем. В гибридных схемах дифференциальная приватность применяется на стороне клиента: перед передачей обновлений серверу локальные градиенты или параметры нормируются и искажаются шумом, что ограничивает возможности извлечения сведений даже при анализе поступающих обновлений центральным сервером. Внедрение DP на стороне клиента требует учета композиции приватности по числу раундов обучения и накладывает дополнительные вычислительные и коммуникационные затраты [5], однако дает формализуемый уровень защиты, сочетающийся с архитектурными преимуществами федеративной схемы.

Для удобства инженерного выбора и системного сравнения практических методов в работе приведена сводная сравнительная Таблица 1 по ключевым критериям.

Таблица 1 – Сравнительная характеристика практических методов обеспечения приватности в мультимодальных генеративных моделях
Table 1 – Comparative characteristics of practical privacy-preserving methods for multimodal generative models

Метод	**Формальные гарантии приватности**	**Пригодность для мультимодальных моделей**	**Накладные вычислительные и коммуникационные расходы**
Дифференциальная приватность (DP)	Формальные гарантии, выражаемые параметрами ε и δ; обеспечивает количефицируемую границу утечек	Применима при явном выборе точки внесения шума; возможно как локальное применение для каждой модальности, так и внесение в общее латентное пространство	Увеличение времени обучения и требований к памяти; потребность в privacy accounting
Федеративное обучение (FL, FedAvg)	Без дополнительного DP формальных гарантий нет; архитектурная защита от передачи исходных данных	Хорошо подходит для распределённых мультимодальных данных, но требует продуманной схемы представления каждого типа данных	Высокие коммуникационные расходы; возможны дополнительные расходы на синхронизацию и управление клиентами
Гибридный подход (DP на клиенте + FL)	При корректной настройке DP на клиентах достигаются формальные гарантии; комбинирует архитектурные и формальные меры	Наиболее пригоден к защите кросс-модальных утечек в распределённых сценариях	Наиболее высокие суммарные расходы: накладные DP + коммуникации FL

Сравнительный анализ показывает, что дифференциальная приватность обеспечивает формальные, количественно измеримые гарантии конфиденциальности и гибко применяется к мультимодальным моделям, но увеличивает вычислительные затраты и требует тщательной настройки параметров. Федеративное обучение снижает прямой риск утечки данных за счет архитектурного подхода, однако не дает формальных гарантий и сопровождается высокими коммуникационными расходами. Гибридный подход, сочетающий DP и FL, обеспечивает наибольшую защиту данных в распределенных мультимодальных системах, сочетая архитектурные и формальные меры, но отличается высокой сложностью интеграции и максимальными ресурсными затратами.

Обсуждение

Ограничения и компромиссы: взгляд практика. Переход от теоретических гарантий к практической реализации механизмов приватности неизбежно сопровождается рядом компромиссов, которые инженер вынужден учитывать при проектировании мультимодальных генеративных систем. Одним из центральных противоречий является компромисс между уровнем приватности и качеством генерации. При использовании дифференциально-приватных методов, таких как DP-SGD, добавление шума в градиенты или латентные представления, снижает способность модели точно воспроизводить структуру данных, поскольку шум в процессе оптимизации приводит к потерям информации, которую модель может эксплуатировать для генерации высококачественного контента. Эта деградация качества наблюдалась как в задачах обучения классификаторов, где сильные настройки приватности (низкие значения ε) существенно ухудшают точность моделей по сравнению с негарантированными аналогами, так и в задачах генерации, где дополнительные шумы могут ухудшать показатели таких метрик качества, как FID или Inception Score, отражающих реалистичность и разнообразие сгенерированных изображений [6].

Вычислительные и коммуникационные затраты составляют еще один аспект практических компромиссов. Дифференциально-приватный стохастический градиентный спуск требует нормализации и обогащения каждого шага оптимизации шумом, что увеличивает время обучения и объем вычислений по сравнению с обычным SGD: нормализация per-example gradient и добавление шумового члена увеличивают сложность каждого шага, а также требуют дополнительных трекинговых вычислений для подсчета бюджета приватности. При применении этих методов в больших мультимодальных архитектурах вычислительные расходы могут быть значительными, особенно если модель имеет сверхпараметризованные энкодеры и глубокие латентные пространства. Схемы федеративного обучения, где обучение проводится распределенно, добавляют к этому коммуникационные расходы, так как на каждом раунде большое количество параметров или обновлений должно передаваться между клиентами и сервером, что в случаях мультимодальных моделей может быть особенно затратным из-за большого объема параметров [7].

Кроме того, механизм приватности может усложнять процесс выравнивания модальностей в мультимодальных системах. Добавление шума в латентные представления или градиенты затрудняет обучение точных и устойчивых взаимосвязей между текстом и изображением или между визуальными и аудиоданными, поскольку шум снижает сигнал, который модель могла бы использовать для согласования семантики различных модальностей. В результате такого вмешательства может ухудшаться согласованность между модальностями: например, модель может генерировать изображение, которое не полностью соответствует текстовому описанию или теряет детали, важные для семантического сопоставления. Это явление усиливается

по мере увеличения уровня шума, необходимого для достижения заданного уровня параметра ε, что отражает реальный инженерный компромисс между строгими гарантиями приватности и качеством мультимодального согласования [8].

Наконец, существующие формальные гарантии приватности имеют свои ограничения в практических условиях. Дифференциальная приватность обеспечивает измеримый уровень защиты информации о присутствии отдельных записей в обучающем наборе, но она не защищает от всех возможных способов утечки знаний модели как таковой. Реальные модели могут по-прежнему представлять общие закономерности данных или статистические особенности, которые могли бы быть использованы злоумышленником для вывода свойств целых подгрупп или характеристик распределения данных. Кроме того, формальные гарантии DP редко учитывают сложные зависимости в больших мультимодальных пространствах, что делает оценку риска утечки через косвенные каналы, такие как реконструкция данных с использованием мощных внешних априорных моделей (например, диффузионных моделей), сложной задачей, выходящей за пределы традиционных аналитических оценок [9].

Инструменты и фреймворки для реализации. Для реализации дифференциальной приватности в обучении нейронных сетей наибольшее распространение получили библиотеки Opacus для PyTorch и TensorFlow Privacy для экосистемы TensorFlow. Эти инструменты предоставляют реализацию дифференциально-приватного стохастического градиентного спуска, включающую нормализацию градиентов отдельных обучающих примеров, добавление контролируемого шума и подсчет накопленного бюджета приватности. В контексте мультимодальных генеративных моделей данные библиотеки позволяют встраивать механизмы приватности как на уровне всей модели, так и локально, например, для отдельных энкодеров модальностей или общего латентного представления.

Ниже, на Рисунке 1, приведен иллюстративный пример интеграции дифференциальной приватности в тренировочный цикл мультимодальной модели.

```
# Инициализация мультимодальной модели
model = MultimodalGenerator()

# Инициализация оптимизатора
optimizer = Optimizer(model.parameters(), learning_rate=η)

# Подключение механизма дифференциальной приватности
privacy_engine = DifferentialPrivacyEngine(
    model=model,
    max_grad_norm=C,        # параметр clipping
    noise_multiplier=σ      # уровень шума
)

privacy_engine.attach(optimizer)

# Тренировочный цикл
for batch in training_data:
    loss = model.compute_loss(batch)
    loss.backward()
    optimizer.step()
    optimizer.zero_grad()

# Оценка накопленного бюджета приватности
epsilon = privacy_engine.get_epsilon(delta=δ)
```

Рисунок 1 – Пример интеграции дифференциальной приватности в тренировочный цикл мультимодальной модели

Figure 1 – Example of integrating differential privacy into the training cycle of a multimodal model

Пример демонстрирует принцип внедрения DP-SGD и не ориентирован на воспроизводимость конкретного эксперимента или использование определенной программной библиотеки.

В данном тренировочном цикле дифференциальная приватность обеспечивается за счет ограничения вклада отдельных обучающих примеров в обновление параметров модели и добавления случайного шума к агрегированным градиентам. Контроль параметров C, σ, а также накопленного значения ε позволяет инженеру управлять компромиссом между качеством генерации и уровнем формальных гарантий конфиденциальности. Если ε определяет допустимый уровень утечки информации, то δ допускает маловероятное событие, при котором утечка может превысить этот уровень. На практике δ задает допустимый риск «срыва» защиты, обычно выбираясь как пренебрежимо малая величина (например, 10^{-5}). Аналогичные схемы могут быть реализованы при внесении шума в латентные представления мультимодальной модели, что в ряде случаев упрощает масштабирование на крупные архитектуры.

Федеративное обучение реализуется с использованием специализированных фреймворков, таких как NVIDIA FLARE, Flower и PySyft, которые обеспечивают координацию распределенных клиентов, агрегацию локальных обновлений и управление процессом обучения [10]. В мультимодальных сценариях эти фреймворки позволяют учитывать гетерогенность данных, когда разные клиенты обладают различными типами модальностей или используют отличающиеся архитектуры локальных моделей. Федеративное обучение снижает риск прямой утечки исходных данных за счет архитектурного принципа распределенного обучения, однако без дополнительных механизмов не предоставляет формальных гарантий приватности на уровне отдельных записей.

Проектирование приватных мультимодальных генеративных моделей требует системного подхода к выбору методов защиты, учитывающего особенности данных, архитектуру модели и требования к формальным гарантиям конфиденциальности. На практике не существует универсального решения, одинаково эффективного для всех сценариев, поэтому выбор стратегии должен основываться на анализе конкретного прикладного контекста и допустимых инженерных компромиссов.

Ключевым фактором при принятии проектных решений является характер доступных данных и организация процесса обучения. В централизованных сценариях, где все мультимодальные данные доступны в едином хранилище и отсутствуют строгие ограничения на их перемещение, основным инструментом обеспечения формальных гарантий приватности выступает дифференциальная приватность, внедряемая непосредственно в процесс обучения модели. В таких условиях применение DP-SGD или механизмов зашумления латентных представлений позволяет формализовать уровень защиты и контролировать утечки информации на уровне отдельных записей, однако сопровождается снижением качества генерации и ростом вычислительных затрат. В распределенных сценариях, где данные физически или юридически не могут быть объединены, предпочтительным архитектурным решением становится федеративное обучение, позволяющее исключить передачу исходных данных за пределы локальных узлов. При этом для достижения формальных гарантий конфиденциальности федеративные схемы целесообразно дополнять механизмами дифференциальной приватности на стороне клиента.

Рекомендации по проектированию и выбору стратегии. Выбор стратегии также определяется архитектурой мультимодальной модели и способом согласования модальностей. Для моделей с независимыми энкодерами и поздней агрегацией модальностей применение механизмов приватности на уровне отдельных энкодеров

позволяет более гибко управлять уровнем шума и минимизировать деградацию качества представлений. В архитектурах с общим латентным пространством и плотным кросс-модальным выравниванием предпочтение может отдаваться внесению шума в общее представление, что обеспечивает защиту кросс-модальных зависимостей, но требует более осторожной настройки параметров приватности. В обоих случаях инженер должен учитывать, что усиление формальных гарантий конфиденциальности неизбежно влияет на способность модели формировать устойчивые и семантически согласованные мультимодальные представления.

С практической точки зрения целесообразно рассматривать гибридные подходы, в которых архитектурные меры защиты сочетаются с формальными механизмами приватности. Такие решения позволяют распределить нагрузку между различными уровнями системы: федеративное обучение снижает риск прямого доступа к данным, а дифференциальная приватность ограничивает утечки информации через параметры модели и передаваемые обновления. Подобная комбинация особенно оправдана в сценариях с повышенными требованиями к конфиденциальности и регуляторному соответствию.

В качестве иллюстрации практического применения описанных рекомендаций можно рассмотреть эскизный кейс проектирования системы генерации медицинских иллюстраций по текстовым описаниям историй болезней. В данном сценарии мультимодальная генеративная модель используется для создания визуальных материалов, сопровождающих клинические описания, при этом обучающие данные включают чувствительные текстовые и визуальные медицинские записи. Часто возникают юридические и этические ограничения, которые исключают возможность централизованного хранения таких данных. Это, в свою очередь, делает распределенную архитектуру обучения предпочтительной.

При проектировании подобной гибридной системы ключевой инженерной задачей становится нахождение оптимального баланса между тремя фундаментальными аспектами: конфиденциальностью данных, качеством итоговой модели и эффективностью процесса обучения. Механизмы дифференциальной приватности, внося случайный шум в градиенты, неизбежно снижают точность модели и могут замедлять сходимость, требуя большего количества федеративных раундов. В свою очередь, федеративное обучение само по себе вносит статистическую неоднородность (non-IID данные у разных клиентов), что усугубляет проблему конвергенции и может привести к смещенной или нестабильной генерации. Поэтому архитектурные решения должны быть тщательно калиброваны. Например, необходимо определить, на каком уровне модели – на уровне энкодеров отдельных модальностей или в общем генеративном модуле – применение дифференциальной приватности будет наиболее эффективным с точки зрения защиты, но наименее разрушительным для семантической согласованности генерируемых иллюстраций. Кроме того, требуется разработать стратегию отбора клиентов для каждого раунда обучения и агрегации их обновлений (например, с использованием адаптивных методов, подобных FedProx), чтобы смягчить эффекты неоднородности данных и минимизировать количество необходимых коммуникационных циклов, каждый из которых расходует приватный бюджет.

Концепция системы предполагает использование федеративного обучения, при котором медицинские учреждения или исследовательские центры выступают в роли клиентов, обучающих локальные экземпляры мультимодальной модели на собственных данных. Локальная модель включает текстовый энкодер для обработки клинических описаний, визуальный энкодер для работы с изображениями и общий генеративный модуль, формирующий согласованные мультимодальные представления. Перед

передачей обновлений на центральный сервер локальные градиенты подвергаются нормализации и зашумлению в соответствии с выбранным механизмом дифференциальной приватности, что позволяет ограничить утечки информации даже при анализе агрегированных обновлений.

Для наглядной иллюстрации описанного подхода на Рисунке 2 представлен эскиз архитектуры системы генерации медицинских иллюстраций на основе мультимодальной генеративной модели, обучаемой в федеративной среде с использованием механизмов дифференциальной приватности.

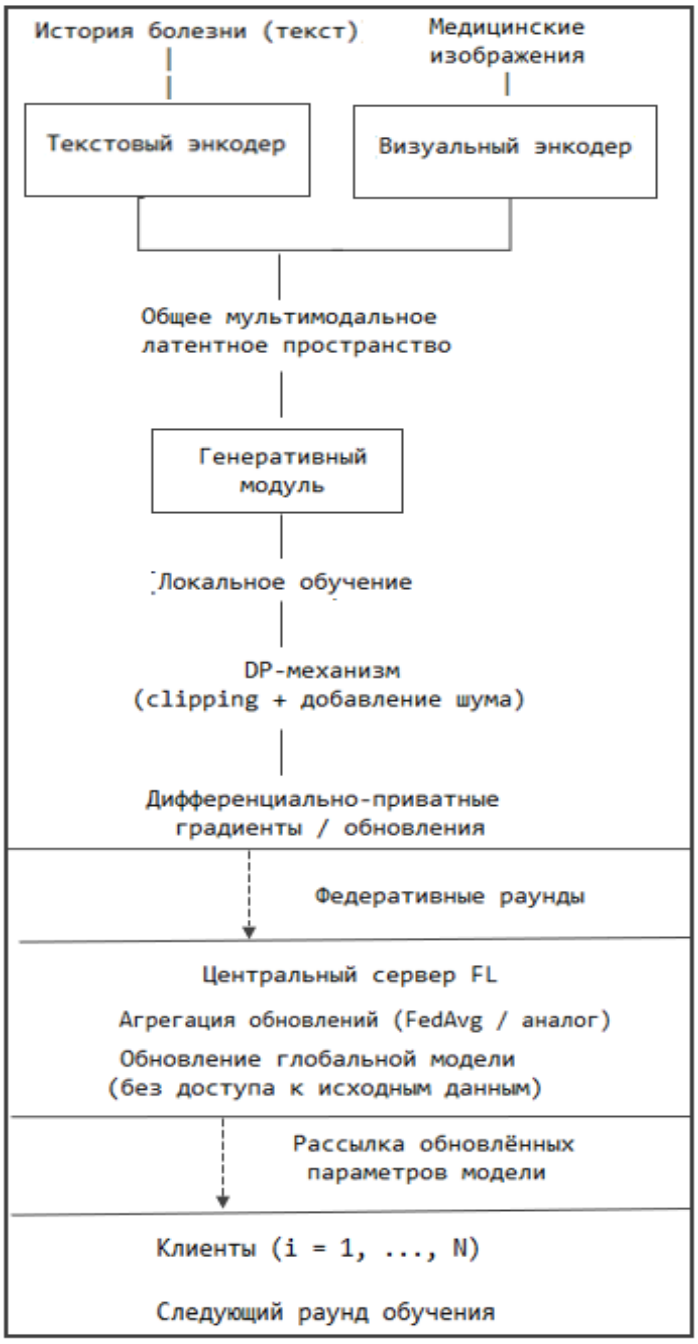

Рисунок 2 – Эскиз архитектуры системы генерации медицинских иллюстраций на основе мультимодальной генеративной модели

Figure 2 – Sketch of the architecture of a medical illustration generation system based on a multimodal generative model

Процесс реализации такой системы может быть разделен на несколько этапов, критически важных для обеспечения как качества генерации, так и конфиденциальности данных. На этапе проектирования архитектуры определяется способ согласования модальностей и точка применения механизмов приватности, что задает базовый компромисс между качеством и уровнем защиты. На этапе обучения осуществляется настройка параметров дифференциальной приватности и контроль накопленного приватного бюджета по мере проведения федеративных раундов. Ключевой точкой контроля является мониторинг деградации качества генерации, позволяющий своевременно скорректировать параметры шума или стратегию обучения. На этапе валидации и развертывания системы проводится оценка устойчивости модели к атакам на приватность и проверка соответствия регуляторным требованиям, что завершает цикл инженерного проектирования.

Концептуальная архитектура системы включает распределенное обучение мультимодальной генеративной модели в федеративной среде. Каждый клиент обучает локальный экземпляр модели на собственных текстовых и визуальных медицинских данных. Для обработки данных используются специализированные энкодеры, формирующие общее мультимодальное латентное представление, на основе которого осуществляется генерация изображений. Перед передачей обновлений на центральный сервер локальные градиенты подвергаются нормализации и зашумлению в соответствии с механизмом дифференциальной приватности. Центральный сервер выполняет агрегацию обновлений без доступа к исходным данным, после чего обновленные параметры модели распространяются на клиентов для следующего раунда обучения.

Заключение

Проведенный анализ показывает, что построение приватных мультимодальных генеративных моделей представляет собой сложную, но в практическом отношении решаемую инженерную задачу. Современные методы обеспечения конфиденциальности, такие как дифференциальная приватность и федеративное обучение, позволяют существенно снизить риск утечек чувствительных данных при обучении и эксплуатации генеративных моделей, однако требуют осознанного выбора архитектурных решений и параметров в зависимости от конкретного прикладного сценария. На практике достижение приемлемого уровня защиты неизбежно связано с компромиссами между качеством генерации, вычислительной сложностью и строгостью формальных гарантий приватности, что делает вопросы проектирования и настройки таких систем ключевыми для инженера.

Перспективы дальнейшего практического развития данной области связаны с созданием более эффективных алгоритмов обучения, устойчивых к зашумлению и масштабируемых для сложных мультимодальных архитектур, развитием специализированных фреймворков, ориентированных на приватное мультимодальное обучение, а также формированием стандартов оценки и аудита конфиденциальности генеративных моделей. Реализация этих направлений позволит повысить надежность и воспроизводимость приватных AI-систем и расширить возможности их безопасного применения в чувствительных и регулируемых областях.

Таким образом, можно сделать следующие выводы, суммирующих основные научно-практические положения работы:

1. Построение приватных мультимодальных генеративных моделей является сложной, но решаемой на практике инженерной задачей.

2. Дифференциальная приватность и федеративное обучение служат технологической основой для минимизации рисков утечек данных на всех этапах жизненного цикла модели.

3. Ключевым фактором успеха является контекстно-зависимое проектирование, требующее тщательного выбора архитектуры и параметров защиты под конкретный сценарий применения.

4. Практическая реализация сопряжена с необходимостью балансировки компромисса между качеством генерации, вычислительными затратами и строгостью гарантий приватности.

5. Основные перспективы развития лежат в области:

– создания масштабируемых и устойчивых к зашумлению алгоритмов обучения;

– разработки специализированных фреймворков для приватного мультимодального обучения;

– формирования стандартов для оценки и аудита конфиденциальности моделей.

6. Развитие указанных исследовательско-инженерных направлений в совокупности создаст основу для повышения надежности приватных AI-систем и расширит возможности их безопасного внедрения в чувствительных и регулируемых областях.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Feretzakis G., Papaspyridis K., Gkoulalas-Divanis A., Verykios V.S. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. *Information*. 2024;15(11). https://doi.org/10.3390/info15110697
2. Naseri M., Hayes J., De Cristofaro E. Local and Central Differential Privacy for Robustness and Privacy in Federated Learning. arXiv. URL: https://arxiv.org/abs/2009.03561 [Accessed 25th November 2025].
3. Sun L., Qian J., Chen X. LDP-FL: Practical Private Aggregation in Federated Learning with Local Differential Privacy. arXiv. URL: https://arxiv.org/abs/2007.15789 [Accessed 25th November 2025].
4. Rafi T.H., Noor F.A., Hussain T., Chae D.-K. Fairness and Privacy-Preserving in Federated Learning: A Survey. arXiv. URL: https://arxiv.org/abs/2306.08402 [Accessed 25th November 2025].
5. Zhu L., Chen X. Privacy protection in federated learning: a study on the combined strategy of local and global differential privacy. *The Journal of Supercomputing*. 2025;81(1). https://doi.org/10.1007/s11227-024-06845-9
6. Катаев А.В., Власова Ю.М., Гусынин Д.А., Ким В.А. Обзор метрик с целью оценки качества работы генеративных моделей для создания изображений. *Инженерный вестник Дона*. 2025;(6). URL: http://www.ivdon.ru/ru/magazine/archive/n6y2025/10119
Kataev A.V., Vlasova Y.M., Gusynin D.A., Kim V.A. A survey of metrics for evaluating the performance of generative models in image creation. *Engineering Journal of Don*. 2025;(6). (In Russ.). URL: http://www.ivdon.ru/en/magazine/archive/n6y2025/10119
7. Рабчевский А.Н. Обзор методов и систем генерации синтетических обучающих данных. *Прикладная математика и вопросы управления*. 2023;(4):6–45.
Rabchevsky A.N. Review of methods and systems for generation of synthetic training data. *Applied Mathematics and Control Sciences*. 2023;(4):6–45. (In Russ.).
8. Xu H., Shrestha Sh., Chen W., Li Zh., Cai Zh. DP-FedLoRA: Privacy-Enhanced Federated Fine-Tuning for On-Device Large Language Models. arXiv. URL: https://arxiv.org/abs/2509.09097 [Accessed 19th December 2025].

9. Ghalebikesabi S., Berrada L., Gowal S., et al. Differentially Private Diffusion Models Generate Useful Synthetic Images. arXiv. URL: https://arxiv.org/abs/2302.13861 [Accessed 19th December 2025].
10. McMahan B., Moore E., Ramage D., Hampson S., Agüera-Arcas B. Communication-Efficient Learning of Deep Networks from Decentralized Data. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), 20–22 April 2017, Fort Lauderdale, FL, USA*. PMLR; 2017. P. 1273–1282.

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Ледовская Екатерина Валерьевна, кандидат технических наук, доцент кафедры прикладной математики, МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: ekvaled@mail.ru

Ekaterina V. Ledovskaya, Candidate of Engineering Sciences, Associate Professor at the Applied Mathematics Department, MIREA – Russian Technological University, Moscow, the Russian Federation.

Статья поступила в редакцию 28.12.2025; одобрена после рецензирования 06.02.2026; принята к публикации 12.02.2026.

The article was submitted 28.12.2025; approved after reviewing 06.02.2026; accepted for publication 12.02.2026.