Федеральное государственное бюджетное учреждение науки
Институт системного программирования им. В. П. Иванникова
Российской академии наук

На правах рукописи

Ушаков Егор Николаевич

Метод условной генерации изображений в диффузионных моделях со структурированным текстовым условием

Специальность 2.3.5 —
«Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
кандидат физико-математических наук
Карпулевич Евгений Андреевич

Москва — 2026

Оглавление

Введение

Актуальность проблемы. Современная прикладная математика и математическая статистика располагают развитым аппаратом для работы с неопределенностью и сложными распределениями: байесовским выводом, методами максимального правдоподобия, вариационными приближениями, стохастической оптимизацией, а также построением параметрических и непараметрических моделей данных. Существенный прогресс последних лет связан с развитием генеративного моделирования, в рамках которого задача формулируется как восстановление (аппроксимация) неизвестного распределения данных и последующая генерация новых примеров из этого распределения.

Значительный вклад в развитие методов анализа изображений, машинного обучения и генеративного моделирования внесли как российские, так и зарубежные научные школы. В России данное направление развивается в рамках школ компьютерного зрения, анализа изображений, обработки больших данных. К ним относятся, в частности, школа Ю. В. Визильтера, связанная с развитием методов обработки и анализа изображений, распознавания визуальных объектов, машинного и технического зрения, а также школа Д. С. Ватолина, ориентированная на методы обработки изображений и видео, компьютерного зрения, машинного обучения, сжатия и оценки качества цифрового видео. Близкие по тематике исследования ведутся также в научных коллективах AIRI, ФИЦ ИУ РАН и Сколтеха, где развиваются методы глубокого обучения, мультимодального анализа данных, генеративного моделирования и прикладного искусственного интеллекта. Отдельно следует отметить направление, связанное со школой И. А. Макарова в AIRI, в рамках которого развиваются методы промышленного искусственного интеллекта, компьютерного зрения, машинного обучения на графах и прикладного анализа данных. В зарубежной научной традиции существенный вклад в становление генеративного моделирования внесли работы Д. Румельхарта, Д. Хинтона и Р. Уильямса по обучению многослойных нейронных сетей, исследования Д. Кингмы и М. Веллинга, предложивших вариационные автокодировщики, И. Гудфеллоу и соавторов, разработавших генеративно-состязательные сети, а

также Я. Сохла-Дикстейна и соавторов, заложивших основы диффузионного подхода к генерации. В дальнейшем Дж. Хо, А. Джейн и П. Аббил развили этот подход в модели диффузионного вероятностного моделирования, а Р. Ромбах и соавторы показали его высокую практическую эффективность в латентном пространстве изображений.

Одной из важных постановок задачи генерации является условная генерация изображений. В этой постановке требуется построить генеративную модель, способную синтезировать объекты данных, в частности изображения, по заданному условию. В качестве такого условия могут выступать текстовое описание, набор признаков, метка класса, контекст или иная дополнительная информация, определяющая требуемые свойства результата. С математической точки зрения здесь возникают две взаимосвязанные задачи: во-первых, необходимо выбрать такое представление условия, которое сохраняет существенную информацию, уменьшает неоднозначность и допускает эффективную вычислительную обработку; во-вторых, необходимо включить это условие в процесс генерации таким образом, чтобы обеспечить не только высокое качество синтезируемых объектов, но и их устойчивое соответствие заданному условию. На практике условие нередко задается в неструктурированном виде - например, в форме текста или набора слабо формализованных признаков, - что затрудняет его интерпретацию моделью и может приводить к снижению управляемости генерации и ухудшению воспроизводимости результатов.

Диффузионные модели представляют собой один из наиболее успешных современных классов генеративных моделей, опирающийся на вероятностную интерпретацию постепенного зашумления данных и обратного стохастического процесса восстановления. Их практическая эффективность обусловлена устойчивостью обучения и способностью аппроксимировать сложные распределения в высоких размерностях. Однако в задаче условной генерации с точки зрения корректности особенно важно, каким образом задано условие и как оно «встраивается» и влияет на обратный процесс: недостаточно получить реалистичные примеры, необходимо обеспечить стабильное выполнение ограничений и уменьшить неопределенность интерпретации условия.

Описанные методы и алгоритмы нашли широкое применение в прикладных областях, где данные имеют сложную структуру, высокую

размерность и существенную внутреннюю вариативность. К ним относятся анализ изображений и видеоданных, обработка текстовой и мультимодальной информации, моделирование сложных объектов и процессов, а также построение интеллектуальных систем поддержки принятия решений. Во всех этих задачах возникает необходимость не только в извлечении закономерностей из имеющихся данных, но и в построении моделей, способных воспроизводить структуру исходного распределения, формировать новые примеры и управляемо изменять их свойства в соответствии с заданными условиями.

С математической точки зрения генерация синтетических данных связана с задачами аппроксимации сложных многомерных распределений, оценки неопределенности, построения устойчивых представлений данных и разработки алгоритмов оптимизации, пригодных для обучения моделей с большим числом параметров. При этом особую роль играет условная генерация, поскольку в прикладных задачах требуется не просто синтезировать реалистичные объекты, а получать данные, удовлетворяющие заданным ограничениям: классовым меткам, текстовым описаниям, наборам признаков, структурным характеристикам или иным формализованным условиям. Поэтому качество генерации определяется не только архитектурой модели, но и тем, насколько полно, компактно и однозначно задано условие.

Практическое применение генеративных моделей часто ограничено дефицитом размеченных данных, дисбалансом классов, редкостью отдельных состояний или конфигураций объектов, а также трудоемкостью получения экспертной разметки. В таких условиях синтетические данные могут использоваться для расширения и балансировки обучающих выборок, моделирования редких случаев, анализа устойчивости алгоритмов и тестирования интеллектуальных систем. Однако использование синтетических данных требует обеспечения их соответствия заданному условию: если условие задано неоднозначно или содержит нерелевантную информацию, модель может воспроизводить формально правдоподобные, но плохо управляемые и слабо воспроизводимые результаты.

На практике условие нередко задается в неструктурированном виде - например, в форме свободного текстового описания, набора слабо формализованных признаков или сообщения эксперта. Такие представления

могут содержать избыточные фрагменты, неоднозначные формулировки, отрицания, ссылки на контекст и признаки, не имеющие непосредственного соответствия в генерируемом объекте. Это затрудняет интерпретацию условия моделью, снижает управляемость генерации и ограничивает возможность формального анализа получаемого результата. В связи с этим актуальной является разработка методов структурированного задания условия диффузионных моделей, основанных на разложении исходного текстового описания на значимые компоненты и представлении его в форме, пригодной для алгоритмической обработки.

В настоящей работе указанные задачи рассматриваются на примере генерации медицинских изображений, где проблема управляемой условной генерации имеет практическую значимость. Медицинские изображения характеризуются высокой размерностью, сложной визуальной структурой, ограниченной доступностью, высокой стоимостью экспертной разметки и необходимостью строгого соответствия синтезируемого изображения заданному клиническому описанию. При этом текстовые описания исследований и врачебные заключения содержат сложную клиническую семантику, допускают вариативность формулировок и представлены в неструктурированном виде. Поэтому применение диффузионных моделей со структурированным заданием условия в данной области позволяет исследовать общий математический и алгоритмический подход к управляемой генерации на практически важном и сложном классе данных.

Целью данной работы является разработка и экспериментальная валидация метода условной генерации синтетических изображений на основе диффузионных моделей со структурированным заданием условия, формируемым из текстового описания, а также создание программной реализации, обеспечивающей повышение качества генерации и практической полезности синтетических данных.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Разработать метод разложения текстового описания для задачи диффузионной генерации изображений, позволяющий выделять семантически значимые компоненты и представлять их в виде структурированного условия, и исследовать его на примере медицинских

данных.

2. Разработать способ использования структурированного условия в латентной диффузионной модели и оценить его влияние на метрики качества генерации по сравнению с использованием неструктурированного текста.
3. Выполнить экспериментальную проверку предложенного метода формирования структурированного условия на медицинских данных, включающую сопоставление с базовым способом задания условия, анализ сходства синтетических и обучающих изображений, оценку возможности определения принадлежности данных к обучающей выборке и риска повторной идентификации пациентов, а также оценку практической полезности синтетических данных в downstream-задачах классификации.

Объектом исследования являются процессы условной генерации изображений на основе диффузионных моделей.

Предметом исследования являются методы и алгоритмы формирования структурированного текстового условия, способы его использования в латентных диффузионных моделях, а также программные средства и экспериментальные методы оценки качества и практической полезности синтетических изображений на примере медицинских данных.

Основные положения, выносимые на защиту:

1. Новый метод разложения текстового условия для диффузионной генерации изображений, обеспечивающий выделение значимых признаков и формирование структурированного условия.
2. Алгоритмы в составе метода разложения текстового условия и аналитические оценки их вычислительной и пространственной сложности через доказательство соответствующих теорем.
3. Архитектура и программная реализация комплекса условной генерации изображений со структурированным текстовым условием, объединяющего методы разложения текстового условия, формирования структурированного условия, обучение и применение латентной диффузионной модели, а также средства комплексной экспериментальной оценки на примере медицинских данных.

Научная новизна. Разработан новый метод структурированного представления текстового условия в задаче диффузионной генерации

изображений, основанный на разложении неструктурированного текстового описания на семантически значимые компоненты, их нормализации и формировании компактного условия генерации; метод разработан и исследован на примере медицинских данных. Разработаны алгоритмы формирования структурированного условия, включающие выделение значимых сущностей и их атрибутов, нормализацию, фильтрацию нерелевантной информации и разрешение противоречий. Для разработанных алгоритмов получены и доказаны оценки вычислительной и пространственной сложности. Показано, что замена исходного текстового условия его структурированным представлением не изменяет асимптотическую сложность обучения и генерации латентной диффузионной модели.

Теоретическая и практическая значимость. Теоретическая значимость работы заключается в развитии подходов к условному генеративному моделированию в диффузионных моделях за счет формализации процесса преобразования неструктурированного текстового описания в структурированное условие и анализа вычислительных свойств входящих в него алгоритмов. Полученные формальные представления, алгоритмическая схема и аналитические оценки могут использоваться как научная и методическая основа в последующих исследованиях и диссертационных работах, посвященных условной генерации, структурированию мультимодальных данных и оценке синтетических изображений. Практическая значимость заключается в программной реализации воспроизводимого конвейера, который может служить основой для разработчиков аналогичных систем: от подготовки текста и формирования условия до обучения и применения латентной диффузионной модели и комплексной оценки синтетических данных. Разработанные решения могут применяться при создании систем расширения и балансировки обучающих выборок, моделирования редких состояний и тестирования алгоритмов анализа медицинских данных.

Апробация работы. Результаты работы докладывались на следующих конференциях:

1. Открытая конференция ИСП РАН, Москва, Россия, декабрь 2021.
2. Открытая конференция ИСП РАН, Москва, Россия, декабрь 2023.
3. IV Открытая конференция молодых ученых ГБУЗ “НПКЦ ДиТ ДЗМ”,

Москва, Россия, апрель 2024.

4. Международная конференция «Иванниковские чтения», Калужская область, Россия, май 2026.

Личный вклад. Все выносимые на защиту результаты получены лично автором.

Публикации. Основные результаты по теме диссертации изложены в четырех работах, опубликованных в изданиях, рекомендованных ВАК.

В статье [1, 2] совместно с соавторами сформулирована задача. Автором выполнена основная часть работы: разработка архитектуры и обучение модели, проведение вычислительных экспериментов, анализ результатов и подготовка к публикации.

В работе [3] автором была разработана постановка задачи исследования. Автор принимал участие в обсуждении методологии и интерпретации результатов.

В статье [4] автором была разработана постановка задачи исследования и план экспериментов. Автор участвовал в постановке и планировании экспериментальных исследований, осуществлял проверку и валидацию полученных результатов, а также выполнял подготовку текста публикации.

Объем и структура работы. Диссертационная работа состоит из введения, трех глав и заключения. Работа изложена на 119 страницах, содержит 9 рисунков и 22 таблицы.

Глава 1. Обзор методов условной генерации изображений и подходов к представлению текстового условия

1.1 Основные подходы к генеративному моделированию

Генеративное моделирование рассматривает задачу восстановления неизвестного распределения данных и последующего получения новых объектов, статистически согласованных с наблюдаемой выборкой. Пусть

$$\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}, \qquad x^{(i)} \sim p_{\text{data}}(x),$$

где $\mathcal{D}$ - обучающая выборка из N объектов, а $p_{\text{data}}(x)$ - неизвестное распределение реальных данных. Требуется построить параметрическую модель $p_\theta(x)$ с параметрами θ, которая приближает $p_{\text{data}}(x)$ и допускает получение новых выборок $\widetilde{x} \sim p_\theta(x)$. Если плотность $p_\theta(x)$ может быть вычислена явно, естественным критерием является максимизация логарифмического правдоподобия:

$$\widehat{\theta}_{\text{ML}} = \arg\max_{\theta} \sum_{i=1}^{N} \log p_\theta\left(x^{(i)}\right),$$

где $\widehat{\theta}_{\text{ML}}$ - оценка параметров методом максимального правдоподобия. Для изображений высокой размерности непосредственное задание такой плотности затруднено: число возможных конфигураций чрезвычайно велико, распределение многомодально, а наблюдаемые изображения сосредоточены в сложной области пространства пикселей. Поэтому современные генеративные модели используют скрытые переменные, авторегрессионное разложение, обратимые преобразования, состязательное обучение либо последовательное зашумление и восстановление данных.

На Рисунке 1 показана общая идея генеративного моделирования: выборка z из простого распределения в латентном пространстве преобразуется генеративной моделью в объект x из сложного распределения данных. В различных классах моделей это преобразование и способ обучения задаются по-разному.

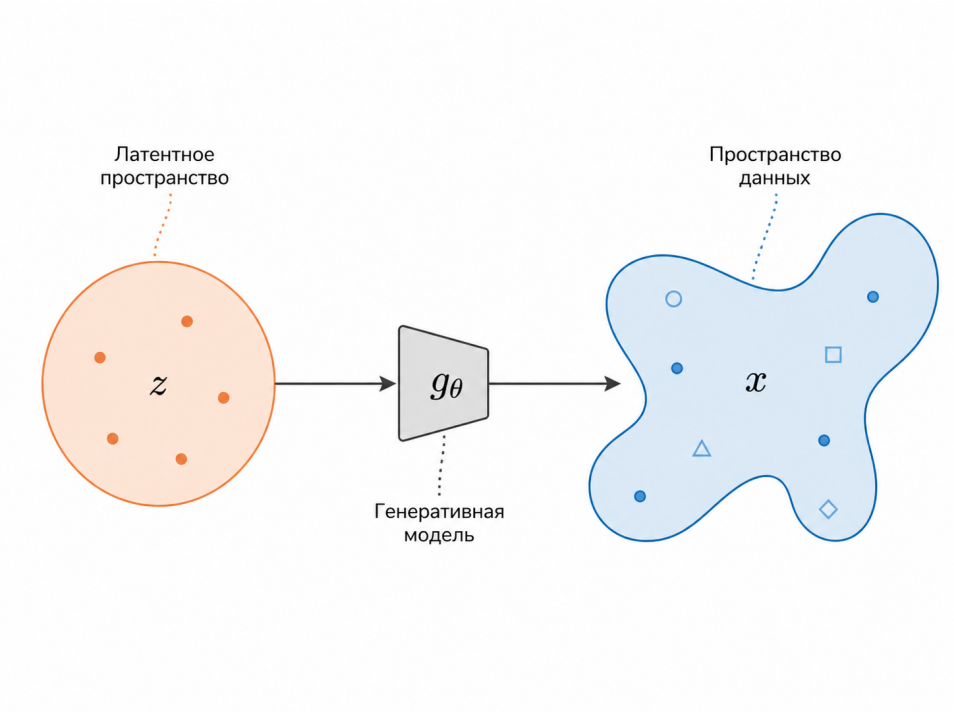


Рисунок 1 — Общая схема генеративного моделирования: преобразование выборки из латентного пространства в объект из распределения данных

1.1.1 Историческое развитие генеративного моделирования

Задача генеративного моделирования возникла не одновременно с современными глубокими нейронными сетями. Её ранние постановки были тесно связаны с вероятностным моделированием: требовалось не только распознать наблюдаемый объект, но и построить модель, способную объяснять структуру данных и порождать новые примеры из того же распределения. Одним из ранних примеров нейросетевой вероятностной генеративной модели стала машина Больцмана. В работе Д. Акли, Д. Хинтона и Т. Сейновски была предложена стохастическая сеть, обучающая внутреннюю энергетическую модель таким образом, чтобы распределение состояний сети приближало распределение наблюдаемых примеров [5]. Этот подход сформировал важную идею: модель должна не только вычислять предсказание, но и описывать вероятностную структуру наблюдений.

В 1990-е годы это направление было развито в моделях с раздельными генеративным и распознающим путями. Алгоритм wake–sleep Д. Хинтона, П. Даяна, Б. Фрея и Р. Нила обучал многослойную стохастическую сеть в двух чередующихся фазах: распознающая часть строила скрытое представление входа, а генеративная часть восстанавливала наблюдение

из скрытых переменных [6]. В этих работах уже присутствует принцип, который сохраняется в современных латентных моделях: вводится скрытое представление, из которого затем восстанавливается наблюдаемый объект.

Новый этап развития глубоких генеративных моделей связан с возобновлением интереса к многослойным вероятностным моделям и автоэнкодерам в середине 2000-х годов. Д. Хинтон, С. Осиндеро и Й.-В. Те предложили эффективное послойное обучение глубоких сетей доверия, а Д. Хинтон и Р. Салахутдинов показали, что глубокие автоэнкодеры могут формировать нелинейные низкоразмерные представления сложных данных [7, 8]. Эти результаты были важны не только для уменьшения размерности: они продемонстрировали практическую возможность обучения глубоких моделей, в которых скрытое пространство содержательно связано со структурой исходных данных.

В 2013–2014 годах вариационные автокодировщики Д. Кингмы и М. Веллинга объединили глубокие нейронные сети с вариационным выводом и репараметризацией, что позволило обучать вероятностные латентные модели градиентными методами в сквозном режиме [9]. Почти одновременно И. Гудфеллоу и соавторы предложили генеративно-состязательные сети, в которых генератор и дискриминатор оптимизируются в рамках состязательной игры [10]. Эти два направления существенно изменили практику генерации изображений: VAE сделали удобным обучение непрерывного латентного пространства, а GAN показали возможность получать визуально детализированные изображения без явного вычисления плотности.

Параллельно развивались модели с точным или явно трактуемым правдоподобием. Авторегрессионные PixelRNN и PixelCNN описывали распределение изображения как произведение условных распределений отдельных пикселей [11, 12], а нормализующие потоки RealNVP и Glow строили цепочку обратимых преобразований, позволяющих вычислять точную плотность и выполнять точный переход между пространством данных и латентным пространством [13, 14].

Состязательные модели также быстро усложнялись. DCGAN показал, что свёрточная архитектура и ряд ограничений на устройство генератора и дискриминатора делают GAN применимыми к устойчивому обучению на изображениях [15]. В WGAN было предложено заменить исходный

состязательный критерий приближением расстояния Вассерштейна, что улучшило поведение градиентов и сделало кривую обучения более содержательной [16]. Архитектура StyleGAN продемонстрировала масштабно-зависимое управление синтезом и высокое качество генерации лиц, разделяя грубые атрибуты и стохастические мелкие детали [17]. Эти работы сделали GAN доминирующим практическим подходом к фотореалистичному синтезу изображений во второй половине 2010-х годов.

Диффузионный подход был предложен Я. Сохлом-Дикстейном и соавторами в 2015 году как вероятностная модель, основанная на последовательном разрушении структуры данных шумом и обучении обратного процесса [18]. Первоначально такие модели уступали наиболее успешным GAN по визуальному качеству и скорости генерации, однако работа Дж. Хо, А. Джейна и П. Аббила показала, что денойзинговая параметризация DDPM позволяет получать высококачественные изображения при стабильном обучении [19]. В дальнейшем было показано, что диффузионные модели могут превосходить GAN по стандартным метрикам синтеза [20], а развитие ускоренных алгоритмов выборки, text-to-image условия и латентной диффузии сделало этот класс практически применимым для генерации изображений высокого разрешения [21, 22, 23, 24].

Таким образом, развитие генеративного моделирования нельзя представлять как простую последовательную замену одного класса моделей другим. Каждый подход решал собственную часть общей проблемы: вероятностные энергетические модели формализовали распределение, автоэнкодеры и VAE развили скрытые представления, GAN улучшили визуальную детализацию, авторегрессионные модели и потоки предоставили явное правдоподобие, а диффузионные модели предложили устойчивую итеративную процедуру генерации и удобный механизм включения условия. В настоящей работе эти направления рассматриваются прежде всего с точки зрения задачи условной генерации изображений со структурированным текстовым условием; разработка и экспериментальная проверка предлагаемого подхода выполняются на примере медицинских данных.

1.1.2 Вариационные, состязательные, авторегрессионные модели и нормализующие потоки

Вариационные автокодировщики. Вариационный автокодировщик (variational autoencoder, VAE) вводит скрытую переменную z с априорным распределением $p(z)$, генеративную модель $p_\theta(x \mid z)$ и приближённое апостериорное распределение $q_\varphi(z \mid x)$ [9]. Для одного объекта x логарифмическое правдоподобие ограничивается снизу вариационной нижней оценкой:

$$\begin{aligned}\log p_\theta(x) \geq \mathcal{L}_{\mathrm{ELBO}}(x) = \mathbb{E}_{q_\varphi(z|x)}\left[\log p_\theta(x \mid z)\right] \\ - D_{\mathrm{KL}}\left(q_\varphi(z \mid x) \,\|\, p(z)\right).\end{aligned} \tag{1.1}$$

Здесь θ - параметры декодера, φ - параметры энкодера, D_{KL} - дивергенция Кульбака–Лейблера. Первый член выражения (1.1) поощряет точное восстановление наблюдения, а второй регуляризует распределение скрытых представлений относительно априорного $p(z)$.

При гауссовском $q_\varphi(z \mid x) = \mathcal{N}(\mu_\varphi(x), \mathrm{diag}(\sigma^2_\varphi(x)))$ стохастическая выборка реализуется репараметризацией

$$z = \mu_\varphi(x) + \sigma_\varphi(x) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I),$$

где $\mu_\varphi(x)$ и $\sigma_\varphi(x)$ - векторы среднего и стандартного отклонения, $\odot$ обозначает поэлементное умножение, а I - единичную ковариационную матрицу. Репараметризация отделяет источник случайности ε от параметров энкодера и позволяет использовать обычное обратное распространение ошибки.

Ключевым достижением VAE стало формирование непрерывного регуляризованного латентного пространства, в котором возможны интерполяция, выборка и управляемое изменение факторов представления. VAE хорошо подходят для задач, где важны латентные признаки и устойчивость оптимизации. Ограничение базовой постановки связано с компромиссом между реконструкцией и регуляризацией: простая пиксельная функция правдоподобия склонна усреднять несколько допустимых вариантов высокочастотных деталей, поэтому изображения могут выглядеть сглаженными. Дополнительной проблемой является вырождение скрытых переменных (posterior collapse), когда мощный декодер частично игнорирует z.

Генеративно-состязательные сети. GAN задаёт распределение неявно: генератор G_θ преобразует шум $z \sim p(z)$ в синтетический объект, а дискриминатор $D_\psi(x)$ стремится различать реальные и сгенерированные примеры [10]. Классическая минимаксная постановка имеет вид

$$\min_\theta \max_\psi \left\{ \mathbb{E}_{x \sim p_{\text{data}}} \left[\log D_\psi(x) \right] + \mathbb{E}_{z \sim p(z)} \left[\log \left(1 - D_\psi(G_\theta(z)) \right) \right] \right\}, \tag{1.2}$$

где θ и ψ - параметры генератора и дискриминатора соответственно. После обучения новый объект генерируется за один прямой проход G_θ, что является существенным преимуществом GAN по скорости выборки.

Практические варианты GAN существенно развили исходную постановку. DCGAN адаптировал состязательное обучение к глубоким свёрточным сетям и показал, что генератор способен обучать содержательные визуальные представления [15]. WGAN изменил геометрию оптимизируемого расхождения и сделал процесс обучения устойчивее [16]. StyleGAN перенёс управление стилем на разные масштабы генератора и достиг высокого качества синтеза изображений лиц [17]. Главным достижением семейства GAN стала возможность получать резкие и визуально правдоподобные изображения при однопроходной генерации. При этом модель не предоставляет прямой оценки правдоподобия, а состязательная оптимизация чувствительна к балансу двух сетей. Коллапс мод может приводить к тому, что разные значения z отображаются в ограниченное множество похожих объектов. Для медицинских данных это особенно нежелательно, поскольку редкие состояния могут быть недостаточно представлены в синтетическом распределении.

Авторегрессионные модели. Авторегрессионный подход задаёт совместное распределение многомерного объекта через цепное правило вероятностей:

$$p_\theta(x) = \prod_{i=1}^{D} p_\theta \left(x_i \mid x_1, \ldots, x_{i-1} \right), \tag{1.3}$$

где D - число последовательно моделируемых компонент объекта, а x_i - i-я компонента в фиксированном порядке. Для изображения компонентой может быть интенсивность отдельного пикселя или канала.

PixelRNN использует двумерные рекуррентные слои для учёта ранее сгенерированных пикселей, а PixelCNN заменяет рекуррентность

маскированными свёртками [11]. Условный PixelCNN показал, что такое распределение можно дополнительно параметризовать меткой класса, признаковым вектором или другим контекстом [12]. Преимуществом авторегрессионных моделей является точное вычисление логарифмического правдоподобия и непосредственное обучение по принципу максимального правдоподобия. Недостаток следует непосредственно из формулы (1.3): при классической выборке компоненты генерируются последовательно, поэтому задержка синтеза возрастает с размерностью изображения. Кроме того, локальная последовательная параметризация не всегда является наиболее удобной для явного управления глобальной структурой изображения.

Нормализующие потоки. Нормализующий поток строит последовательность обратимых преобразований между объектом x и латентной переменной z. Пусть $z = f_\theta(x)$, где f_θ - взаимно однозначное дифференцируемое преобразование, а z имеет простую плотность $p_Z(z)$. Тогда по формуле замены переменных

$$\log p_X(x) = \log p_Z(f_\theta(x)) + \log \left| \det \frac{\partial f_\theta(x)}{\partial x} \right|, \tag{1.4}$$

где p_X - плотность модели в пространстве данных, а определитель якобиана учитывает локальное изменение объёма при преобразовании.

RealNVP использует аффинные coupling-слои с просто вычисляемым якобианом [13]. Glow развивает эту идею с помощью обратимых 1×1-свёрток и демонстрирует возможность генерации и манипулирования изображениями большого размера при сохранении точного правдоподобия [14]. Преимуществами потоков являются точный логарифм правдоподобия, точный переход в латентное пространство и обратимость. Цена этих свойств - архитектурные ограничения: каждый блок должен быть обратимым, а определитель якобиана должен вычисляться достаточно эффективно. При сложных изображениях это ограничивает свободу проектирования сети и может увеличивать потребление памяти. Сравнение основных классов генеративных моделей приведено в Таблице 1.

Практический выбор класса генеративной модели определяется не одной метрикой, а совокупностью требований. Для медицинской генерации важны устойчивость обучения, способность представлять редкие режимы

Таблица 1 — Сравнение основных классов генеративных моделей

Класс	**Принцип обучения**	**Основные достоинства**	**Основные ограничения**
VAE	Вариационная нижняя оценка	Устойчивое обучение, непрерывное латентное пространство, явный энкодер	Компромисс реконструкции и регуляризации, сглаживание деталей, возможное вырождение латентных переменных
GAN	Состязательная оптимизация генератора и дискриминатора	Высокая визуальная детализация, быстрая однопроходная генерация	Нестабильность оптимизации, отсутствие прямого правдоподобия, коллапс мод
Авторегрессионная модель	Максимизация точного условного правдоподобия	Точное правдоподобие, естественное условное моделирование	Последовательная и потому медленная генерация большого числа компонент
Нормализующий поток	Максимизация точного правдоподобия через обратимые преобразования	Точная плотность, точный латентный вывод, обратимость	Жёсткие требования к обратимости и вычислимости якобиана
Диффузионная модель	Денойзинговый критерий для обратного процесса	Устойчивое обучение, хорошее покрытие распределения, гибкое кондиционирование	Итеративная генерация и высокая стоимость многократных проходов денойзера

распределения, возможность включать несколько видов условий и отсутствие необходимости в точном пиксельном соответствии одному эталонному изображению. GAN обладают преимуществом быстрой генерации, однако их состязательный критерий может быть чувствителен к настройке оптимизации и неполному покрытию распределения. VAE дают удобное латентное пространство, но исходная вероятностная постановка содержит компромисс между качеством реконструкции и регулярностью скрытого распределения. Нормализующие потоки обеспечивают обратимость и точную плотность, но достигают этого за счёт архитектурных ограничений. Авторегрессионные модели математически прозрачны с точки зрения правдоподобия, однако последовательная выборка затрудняет генерацию больших изображений. Диффузионные модели снимают часть этих ограничений ценой более дорогой итеративной генерации.

Для рассматриваемой задачи особенно существенны устойчивость обучения, способность покрывать сложное распределение изображений и возможность передавать текстовое условие. В настоящей работе модели сравниваются по количественным метрикам качества генерации, downstream-показателям и результатам анализа конфиденциальности. Быстрая однопроходная генерация также уступает по приоритету устойчивому обучению и возможности управляемого задания условия. Поэтому основой выбран класс диффузионных моделей.

Развитие диффузионных моделей показало, что их математическое описание может быть отделено от конкретной дискретизации процесса. В непрерывном времени прямое зашумление может быть представлено стохастическим дифференциальным уравнением, а генерация - соответствующим обратным SDE или эквивалентным вероятностным ODE [25]. Такой взгляд объединяет несколько вариантов score-based и diffusion-моделей и подчёркивает, что структурирование условия не изменяет вероятностную основу диффузии, а модифицирует дополнительную информацию, с которой оценивается обратная динамика.

1.1.3 Диффузионные модели в пространстве изображений

Диффузионная модель задаёт прямой процесс постепенного зашумления данных и обучаемый обратный процесс восстановления [18, 19]. Пусть $x_0 \sim$

$p_{\text{data}}(x)$ - исходное изображение, $t \in \{1, \ldots, T\}$ - номер шага диффузии, а T - полное число шагов. На каждом шаге марковской цепи к изображению добавляется гауссовский шум, доля которого определяется параметром $\beta_t \in (0, 1)$. Введём $\alpha_t = 1 - \beta_t$ и $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Тогда состояние на произвольном шаге может быть получено непосредственно из x_0:

$$x_t = \sqrt{\overline{\alpha}_t}\, x_0 + \sqrt{1 - \overline{\alpha}_t}\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{1.5}$$

Здесь ε - независимый стандартный гауссовский шум той же размерности, что и x_0, а I - единичная ковариационная матрица. При $\overline{\alpha}_T \approx 0$ распределение x_T близко к стандартному гауссовскому шуму.

Обратный процесс последовательно аппроксимирует переходы $p_\theta(x_{t-1} \mid x_t)$ нейронной сетью, начиная со стандартного гауссовского шума. В DDPM сеть обычно предсказывает добавленный шум $\varepsilon_\theta(x_t, t)$ и обучается по критерию

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{x_0, t, \varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t}\varepsilon, t\right)\right\|_2^2\right]. \tag{1.6}$$

В выражении (1.6) математическое ожидание берётся по $x_0 \sim p_{\text{data}}$, случайно выбранному шагу $t \sim \text{Unif}\{1, \ldots, T\}$ и $\varepsilon \sim \mathcal{N}(0, I)$; ε_θ обозначает нейронную сеть с параметрами θ, предсказывающую добавленный шум. Таким образом, обучение сводится к регрессии к известному шуму и не требует одновременной оптимизации конкурирующих сетей.

Апостериорное распределение прямого процесса $q(x_{t-1} \mid x_t, x_0)$ имеет аналитическую гауссовскую форму, а по предсказанному шуму восстанавливается оценка исходного изображения. Поэтому предсказание шума эквивалентно оценке среднего обратного перехода и связано с оцениванием функции счёта зашумлённого распределения. Это объясняет, почему последовательное применение денойзера перемещает выборку из области гауссовского шума к областям высокой плотности распределения изображений.

Расписание $\{\beta_t\}$ определяет отношение сигнала к шуму на разных этапах. Слишком быстрое зашумление усложняет восстановление, а слишком медленное увеличивает число необходимых шагов. На практике применяются линейные, косинусные и другие расписания, а алгоритм выборки может использовать подмножество временных шагов. Для корректного сравнения

моделей расписание и число шагов генерации должны быть одинаковыми.

Генерация начинается с $x_T \sim \mathcal{N}(0, I)$. На каждом шаге сеть оценивает шум и формирует менее зашумлённое состояние. В качестве денойзера обычно используется U-Net с остаточными блоками, многомасштабными признаками и временными вложениями [26]. Один проход сети обновляет всё изображение, но переходы по времени выполняются последовательно. Методы ускоренной выборки, например DDIM, уменьшают число шагов без изменения основного обучающего критерия [21].

Пиксельная диффузия обеспечивает высокое качество, однако многократная обработка изображения исходного разрешения требует значительных вычислительных ресурсов. Стоимость генерации пропорциональна числу шагов T и стоимости одного прохода U-Net над изображением разрешения $H \times W$. Уменьшение T ускоряет выборку, но при слишком редкой временной сетке может снижать качество восстановления. Это ограничение особенно существенно для медицинских изображений высокого разрешения, где уменьшение H и W способно удалить малые диагностические признаки.

Связь DDPM с оцениванием функции счёта позволяет рассматривать множество алгоритмов генерации как различные численные способы интегрирования одной обратной динамики. В непрерывной формулировке score-функция $\nabla_x \log p_t(x)$ характеризует направление роста логарифма плотности зашумлённых данных, а нейронная сеть аппроксимирует эту величину для различных уровней шума [25]. Из этого следует важное практическое свойство: обученная модель не обязана использовать при генерации ту же дискретную сетку, которая применялась при обучении. Поэтому число шагов и конкретный scheduler являются параметрами процедуры выборки и должны фиксироваться при сравнении вариантов условия.

Большое число последовательных вызовов денойзера долгое время оставалось одним из основных ограничений диффузионных моделей. Помимо DDIM были предложены специализированные численные методы решения, использующие структуру диффузионного ODE. Например, DPM-Solver строит высокопорядковое приближение и позволяет существенно уменьшать число вычислений нейронной сети без повторного обучения модели [27]. Работа Karras

и соавторов показывает, что расписание уровней шума, параметризация сети, предобусловливание и численный solver следует рассматривать как связанные, но разделимые элементы пространства проектных решений [28]. Поэтому корректное экспериментальное сравнение двух способов задания условия требует неизменности не только архитектуры U-Net, но и всей процедуры выборки.

Ещё один аспект связан с параметризацией целевой величины. Сеть может предсказывать непосредственно шум, исходное изображение или линейную комбинацию этих представлений. Различные параметризации могут менять устойчивость оптимизации при разных отношениях сигнала к шуму, однако не меняют принципа условной генерации: контекст должен влиять на оценку обратного перехода на каждом временном шаге. В настоящей работе форма параметризации не является предметом изменения, что позволяет изолировать исследуемый фактор - представление радиологического текста.

Таким образом, диффузионная модель включает два уровня решений. Первый уровень определяет обучаемое вероятностное представление данных и архитектуру денойзера; второй - способ численного прохождения обратного процесса. При исследовании структурированного текстового условия оба уровня должны сохраняться одинаковыми для базовой и предлагаемой моделей. Это требование далее используется при формировании экспериментального протокола третьей главы.

1.1.4 Латентные диффузионные модели и Stable Diffusion

Латентная диффузионная модель переносит процесс зашумления и восстановления из пространства пикселей в сжатое пространство автоэнкодера [24]. Кодировщик $\mathcal{E}_\phi$ и декодировщик $\mathcal{D}_\omega$ задают преобразования

$$z_0 = \mathcal{E}_\phi(x), \qquad \widehat{x} = \mathcal{D}_\omega(z_0),$$

где ϕ и ω - параметры кодировщика и декодировщика, $\widehat{x}$ - реконструкция изображения, а $z_0 \in \mathbb{R}^{H_z \times W_z \times C_z}$ - латентный тензор. Обычно $H_z < H$ и $W_z < W$, где H и W - высота и ширина исходного изображения, а C_z - число каналов латентного представления. Такое обозначение позволяет не смешивать число каналов с символом c, который далее используется для условия генерации. Диффузионный процесс строится по той же схеме, что и формула (1.5), с

заменой x_0 на z_0. Критерий обучения соответствует выражению (1.6).

При наличии условия c сеть дополнительно получает его представление $\tau_\eta(c)$, где τ_η - энкодер условия с параметрами η:

$$\mathcal{L}_{\text{LDM}}^{\text{cond}} = \mathbb{E}_{x,t,\varepsilon,c}\left[\left\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau_\eta(c)\right)\right\|_2^2\right]. \tag{1.7}$$

Здесь z_t - зашумлённое латентное представление изображения на шаге t, а ε_θ - денойзер с параметрами θ. Математическое ожидание вычисляется по обучающим парам (x, c), случайному шагу t и гауссовскому шуму ε. Таким образом, прямой процесс и целевой шум остаются прежними; условие изменяет оценку обратного перехода.

Архитектура Stable Diffusion включает автоэнкодер, U-Net-денойзер, текстовый энкодер и расписание шума. Автоэнкодер должен удалять преимущественно перцептивно избыточную информацию, сохраняя геометрию и признаки, необходимые для восстановления. После завершения обратного процесса итоговый латентный код декодируется в изображение один раз. Если пространственное разрешение уменьшается в f раз по каждой координате, число позиций латентного тензора сокращается приблизительно в f^2 раз. Точное ускорение зависит от числа каналов и архитектуры U-Net, но основная последовательность денойзинга становится существенно дешевле обработки пикселей исходного разрешения.

Текстовое или иное условие кодируется отдельным энкодером и передаётся в U-Net через механизмы перекрёстного внимания. U-Net обрабатывает латентный тензор на нескольких масштабах: высокое разрешение необходимо для локальных деталей, а низкое - для глобальной структуры. Пропускные соединения сохраняют пространственную информацию, а временное вложение сообщает текущий уровень шума. Условные блоки располагаются на нескольких уровнях, поэтому текст может влиять как на композицию, так и на локальные элементы.

Латентная диффузия объединяет устойчивость обучения расшумлением, вычислительную эффективность и универсальный интерфейс условной генерации. При этом автоэнкодер вводит дополнительный источник потерь: клинически важный малый признак должен сохраниться после кодирования и декодирования. Поэтому выбор разрешения латентного пространства является компромиссом между стоимостью вычислений и детализацией. Эти свойства

и ограничения обосновывают выбор Stable Diffusion в качестве основы разрабатываемого метода.

Общая схема условной латентной диффузии, включающая переход в латентное пространство, U-Net и передачу условия через cross-attention, приведена на Рисунке 2.

Латентное представление не предписывает единственную архитектуру денойзера. Первоначально латентные диффузионные модели широко использовали U-Net, однако дальнейшие исследования показали возможность замены свёрточного денойзера трансформерной архитектурой. В Diffusion Transformer (DiT) латентное изображение разбивается на последовательность патчей, которая обрабатывается Transformer; увеличение вычислительного масштаба такой архитектуры сопровождалось улучшением качества генерации на стандартных задачах [29]. Для темы настоящей работы этот результат важен тем, что структурированное условие не привязано к конкретному типу денойзера: оно формируется до генеративной модели и может быть передано в любую архитектуру, поддерживающую текстовый контекст. Stable Diffusion выбран как воспроизводимая базовая реализация латентной диффузии, а не как ограничение области применимости предложенного метода.

Сжатие в латентное пространство одновременно является преимуществом и ограничением. Оно резко сокращает число пространственных позиций, на которых выполняются многократные шаги денойзинга, но переносит часть ответственности за сохранение клинических деталей на автоэнкодер. Малые затемнения, тонкие линии, контуры плевральной полости или положение медицинских устройств могут занимать небольшую долю исходного изображения. Если такие признаки плохо представлены в латентном коде, увеличение качества текстового условия не способно полностью компенсировать потерю визуальной информации. Поэтому оценка медицинской генерации должна учитывать как текстовую управляемость, так и ограничения визуального представления.

С другой стороны, латентная диффузия удобна для экспериментального исследования именно потому, что разделяет три функции: автоэнкодер отвечает за визуальное сжатие, текстовый энкодер - за представление условия, а денойзер - за итеративное восстановление латентного изображения. Изменяя только текст, поступающий в один и тот же энкодер, можно исследовать

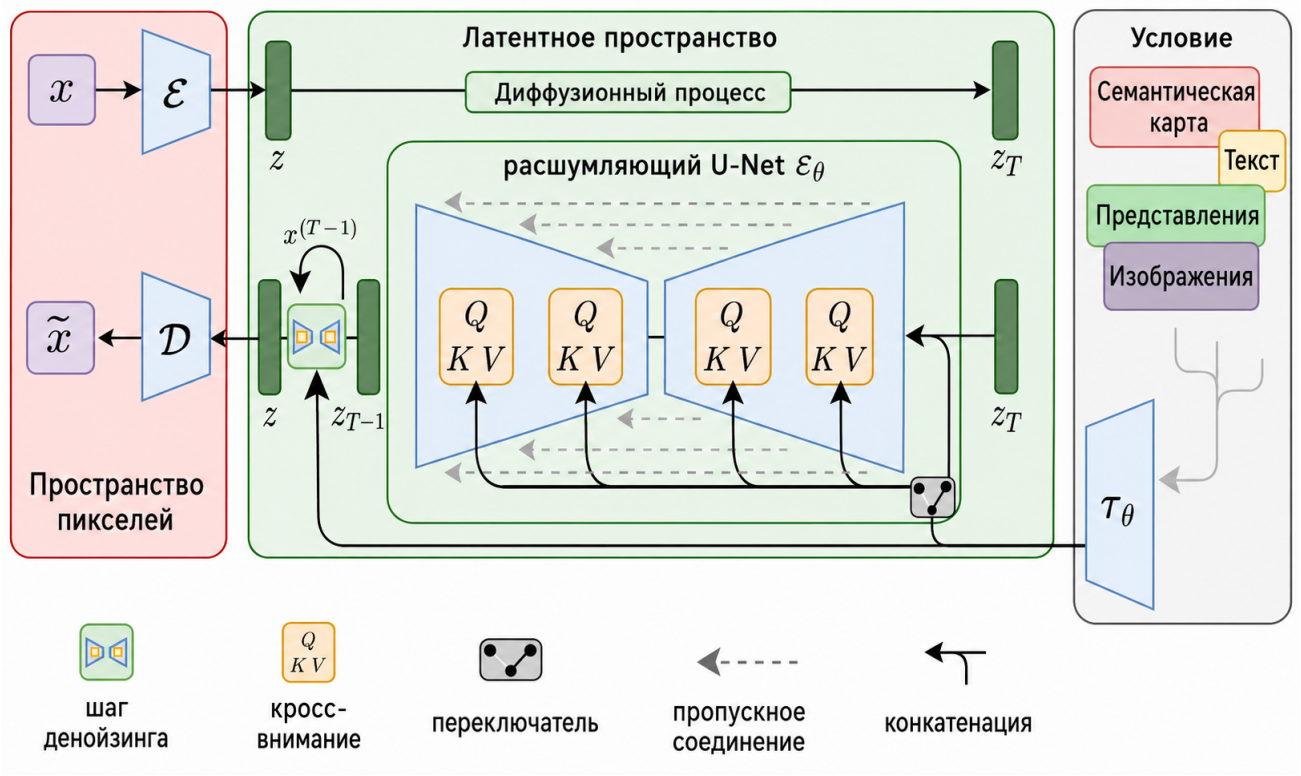


Рисунок 2 — Общая схема условной генерации изображения в модели латентной диффузии

информационную сторону условной генерации без изменения пространства изображений и основных вычислительных блоков. Эта модульность является одной из причин выбора Stable Diffusion для проверки гипотезы о преимуществах структурированного задания радиологического условия.

1.2 Условная генерация изображений

В безусловной постановке модель приближает распределение $p(x)$. В условной генерации требуется моделировать распределение $p_\theta(x \mid c)$, где c задаёт требуемые свойства результата. Для латентной диффузионной модели соответствующий критерий приведён в формуле (1.7).

Архитектура обратного процесса при этом может оставаться неизменной, а управление результатом достигается за счёт способа представления и передачи c.

Качество условной генерации включает два требования. Синтетическое изображение должно принадлежать распределению реалистичных объектов и одновременно соответствовать заданному условию. Усиление влияния c может повысить соответствие, но уменьшить разнообразие или привести к

артефактам. Поэтому выбор формы условия и способа его передачи являются самостоятельной частью постановки задачи.

Условие также выступает информационным ограничением. Если два существенно разных состояния кодируются одинаковым c, генератор вынужден моделировать их смесь. Напротив, избыточное условие с большим числом нерелевантных компонентов усложняет выделение причинно связанных признаков. Желательно, чтобы представление было достаточно полным для различения требуемых состояний, но не содержало контекст, который не должен отражаться в результате.

Развитие text-to-image диффузии в общем домене показало, что качество генерации существенно зависит от взаимодействия диффузионного денойзера и языкового представления. GLIDE продемонстрировал возможности текстово-управляемой диффузии для генерации и редактирования изображений [22], а Imagen показал эффективность использования мощного предобученного языкового энкодера в каскадной text-to-image системе [23]. Эти работы подчёркивают, что текстовый канал является самостоятельной частью генеративного решения, а не только вспомогательным способом выбора класса. Архитектурной основой современных языковых энкодеров служит механизм self-attention Transformer [30], позволяющий формировать контекстные представления токенов с учётом всей входной последовательности. Однако контекстное кодирование не устраняет автоматически специфические для радиологии отрицания, исторические ссылки и атрибутивные связи, что мотивирует предварительное структурирование условия.

1.2.1 Основные способы задания условия

Условие может иметь различную структуру в зависимости от доступной разметки и требуемой управляемости. Независимо от формы оно должно быть преобразовано в представление, совместимое с денойзером. На практике применяются три основных механизма: конкатенация с входными или промежуточными признаками, аффинная модуляция каналов и внимание. При конкатенации условный тензор дополняет визуальные признаки по канальному измерению. Аффинная модуляция масштабирует и сдвигает каналы признаков с параметрами, вычисляемыми из условия. Такой механизм используется, в частности, в FiLM-подобных слоях [31]. Внимание предпочтительно для

последовательностей переменной длины и поэтому широко применяется для текста.

Категориальные метки. Наиболее простой вариант - дискретная метка класса. Её индекс преобразуется в обучаемое векторное представление и добавляется к временным либо промежуточным признакам сети. Такой подход удобен для генерации объектов фиксированного числа классов, но не позволяет подробно задавать локализацию, степень выраженности или сочетание нескольких свойств.

Векторы признаков. Условие может быть представлено вектором $c = (c_1, \ldots, c_K)$, содержащим бинарные, категориальные или вещественные атрибуты. В медицинской задаче это могут быть метки патологий, тип проекции, пол пациента или характеристики исследования. Векторное условие компактно и легко контролируется, однако требует заранее фиксированного словаря признаков и не сохраняет свободную структуру исходного описания.

При многометочной генерации один объект может иметь несколько положительных компонентов. Простая сумма вложений не отражает отношения между ними: локализация и степень выраженности остаются отдельными индексами либо теряются. Расширение вектора всеми возможными сочетаниями приводит к разреженному пространству и недостатку примеров. Поэтому для сложных клинических описаний предпочтительны последовательные или структурированные представления переменной длины.

Пространственные условия. Маска сегментации, карта глубины, контур или разметка ключевых точек задают не только наличие объекта, но и его положение. Такие условия передаются через дополнительные каналы, отдельный энкодер или блоки управления. Пространственная форма особенно полезна для локального редактирования и генерации с заданной анатомией, но требует дорогостоящей пиксельной разметки. Кроме того, слишком точная маска может фактически задавать форму результата и ограничивать вариативность, тогда как грубая область интереса оставляет генератору свободу внутри клинически допустимого диапазона.

Изображение-пример. В качестве условия может использоваться другое изображение: исходный кадр для редактирования, изображение иной модальности или референс стиля. Модель получает его признаки через

энкодер и формирует результат, согласованный с визуальным контекстом. Такой подход обеспечивает детальное управление, но может сохранять индивидуальные особенности исходного объекта и поэтому требует отдельного анализа конфиденциальности.

Текст. Текстовое условие позволяет задавать переменное число объектов и атрибутов без фиксированного набора классов. Оно кодируется языковой моделью и передаётся в диффузионную сеть через механизмы внимания. Текст удобен для медицинских данных, поскольку радиологические отчёты уже содержат описания изображений. Основная проблема состоит в неоднозначности свободного языка: одинаковое состояние может быть описано разными фразами, а часть текста не иметь прямого визуального соответствия.

Мультимодальное условие. Несколько источников информации могут использоваться совместно, например текст, маска и категориальная метка. Это повышает управляемость, но усложняет согласование модальностей и подготовку данных. При противоречии условий модель может непредсказуемо отдавать приоритет одному из них.

Сравнение основных вариантов приведено в Таблице 2.

Таблица 2 — Основные способы задания условия

Условие	**Форма**	**Преимущества**	**Ограничения**
Метка класса	Дискретный индекс	Простота и компактность	Фиксированный набор классов
Вектор признаков	Набор атрибутов	Явное управление несколькими свойствами	Требует заранее заданной схемы
Маска или карта	Пространственный тензор	Точный контроль локализации	Необходима пиксельная разметка
Изображение	Признаки энкодера	Сохранение визуального контекста	Риск копирования индивидуальных особенностей
Текст	Последовательность токенов	Гибкость и переменное число признаков	Неоднозначность и нерелевантные фрагменты

Для настоящей работы ключевым является текстовое условие. Оно позволяет использовать существующие пары “медицинское изображение - радиологический отчёт”, но требует преобразования клинического описания в форму, более однозначную для генеративной модели.

Условное управление может выполняться не только за счёт подачи контекста в нейронную сеть, но и за счёт изменения оценки обратной динамики. В classifier-free guidance условный и безусловный прогнозы одной генеративной модели комбинируются с коэффициентом guidance, позволяющим управлять компромиссом между соответствием условию и разнообразием выборок [32]. При слишком высоком коэффициенте модель может сильнее следовать наиболее выраженным признакам условия, но одновременно снижать разнообразие или усиливать артефакты. Поэтому одинаковое значение CFG является обязательной частью контролируемого сравнения двух текстовых представлений.

Для условий с явной пространственной структурой были разработаны дополнительные архитектурные механизмы. ControlNet вводит обучаемую ветвь для передачи карт границ, глубины, позы, сегментации и других пространственных сигналов в предобученную text-to-image модель [33]. T2I-Adapter решает близкую задачу через относительно компактные адаптеры, связывающие внешние условия с внутренними признаками диффузионной сети [34]. Эти подходы показывают, что понятие условия в диффузионной генерации шире текстового промпта: оно может иметь дискретную, векторную, текстовую или пространственную структуру.

Однако использование пространственной маски или дополнительного изображения требует соответствующей разметки на уровне каждого объекта. Радиологические отчёты, напротив, уже сопровождают значительную часть медицинских изображений и не требуют ручного построения масок специально для генеративной модели. Поэтому в рассматриваемой постановке целесообразно сохранить текстовый интерфейс, но сделать внутреннюю структуру текста более явной. Список находок занимает промежуточное положение между свободным естественным языком и жёстким вектором признаков: число элементов остаётся переменным, но каждый элемент имеет фиксированную семантическую схему.

Именно это различие важно для последующей формализации. Метод не добавляет новый канал conditioning в U-Net и не требует обучения отдельной управляющей сети. Вместо этого преобразуется информационное содержание существующего текстового канала. Благодаря этому результаты можно интерпретировать как эффект структурирования условия, а не как

следствие увеличения числа параметров или добавления пространственной разметки.

1.2.2 Текстовое условие и механизмы его передачи в модель

Пусть текстовое условие c после токенизации представлено последовательностью $(w_1, \ldots, w_M)$. Токенизация может разбивать редкий медицинский термин на несколько подслов, что увеличивает длину и затрудняет сопоставление с одним визуальным понятием. Текстовый энкодер τ_η формирует матрицу контекстных признаков

$$Y = \tau_\eta(c) \in \mathbb{R}^{M \times d_c}.$$

Здесь M - число токенов во входной последовательности, d_c - размерность контекстного представления одного токена, а η - параметры текстового энкодера. В универсальных text-to-image моделях могут использоваться энкодеры, предобученные на больших текстовых или мультимодальных корпусах, например CLIP [35]. Для специализированной медицинской терминологии качество представлений зависит от того, насколько соответствующие слова и сочетания были представлены при предобучении.

Основным механизмом передачи текста в U-Net является перекрёстное внимание. Для промежуточных визуальных признаков $\Phi(z_t)$ вычисляются запросы, а из текстового представления - ключи и значения:

$$Q = \Phi(z_t)W_Q, \qquad K = YW_K, \;\; V = YW_V,$$
$$\mathrm{CrossAttn}(\Phi, Y) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_a}}\right)V.$$

Здесь Q, K и V - матрицы запросов, ключей и значений; W_Q, W_K и W_V - обучаемые матрицы линейных проекций, $\Phi(z_t)$ - промежуточное визуальное представление латента, а d_a - размерность ключа, используемая для масштабирования скалярных произведений. Тем самым каждая пространственная позиция латентного представления получает взвешенную комбинацию текстовых признаков. Вес токена зависит не только от его клинической значимости, но и от сходства ключа с текущим визуальным запросом. Наличие большого числа шаблонных токенов расширяет знаменатель softmax и может рассеивать внимание. Кроме того, cross-attention

не гарантирует логически корректной обработки отрицания: токены названия патологии и маркера “no” представлены отдельно и должны быть объединены контекстным энкодером.

Условие влияет на процесс расшумления на нескольких масштабах U-Net и на каждом шаге обратного процесса. На ранних шагах преимущественно формируется глобальная структура, а на поздних - детали. Поэтому ошибка в условии может проявиться как в общей композиции, так и в локальной патологии.

Другой способ управления - guidance с внешним классификатором. Градиент $\nabla_{x_t} \log p_\phi(c \mid x_t)$ корректирует направление обратного процесса в сторону заданного класса [20]. Этот подход требует отдельного классификатора для зашумлённых состояний и неудобен для свободного текста.

Для усиления следования тексту чаще используется classifier-free guidance [32]. Во время обучения часть условий заменяется пустым, благодаря чему одна сеть оценивает условное и безусловное направления. При генерации используется

$$\widehat{\varepsilon}_\theta = \varepsilon_\theta(z_t, t, \varnothing) + s\left[\varepsilon_\theta(z_t, t, c) - \varepsilon_\theta(z_t, t, \varnothing)\right],$$

где s - коэффициент guidance, а $\varnothing$ обозначает пустое условие, используемое для оценки безусловного направления. При $s > 1$ влияние условия усиливается. Слишком большое значение может снижать разнообразие и создавать неестественные детали, поэтому оно выбирается экспериментально.

Вероятность замены условия на пустое во время обучения также влияет на результат. Если безусловных примеров слишком мало, оценка $\varepsilon_\theta(z_t, t, \varnothing)$ становится неточной; если их слишком много, модель слабее использует текст. Для корректного сравнения исходного и структурированного условия эта вероятность, коэффициент guidance и число шагов должны быть одинаковыми.

В некоторых text-to-image системах дополнительно используется отрицательное условие, перечисляющее нежелательные свойства. В медицинской задаче такое представление следует отличать от отрицательных клинических утверждений. Фраза об отсутствии патологии описывает текущее состояние изображения, тогда как negative prompt задаёт направление, от которого алгоритм выборки должен удаляться. Непосредственная подстановка одного вместо другого не является семантически эквивалентной.

Текстовое условие имеет несколько ограничений. Во-первых, энкодер

работает с последовательностью ограниченной длины; длинные отчёты могут усекаться. Во-вторых, все токены участвуют в механизме внимания, включая шаблонные и нерелевантные фрагменты. В-третьих, отрицание, неопределённость и связь признака с анатомической областью должны быть восстановлены из контекста. Наконец, разные формулировки одного состояния могут давать отличающиеся векторные представления.

Для общего text-to-image корпуса эти ограничения частично компенсируются масштабом данных. В медицинском домене число пар меньше, терминология специализирована, а цена семантической ошибки выше. Токен, обозначающий патологию, может встречаться и в положительном, и в отрицательном контексте; сравнительные конструкции зависят от недоступного предыдущего исследования; редкие термины представлены недостаточно. Следовательно, проблема не сводится к увеличению коэффициента guidance: усиление неоднозначного условия лишь усиливает его неверную интерпретацию.

Структурирование текста должно уменьшать число нерелевантных токенов и явно связывать находку с её атрибутами. При этом сохранение текстового интерфейса позволяет использовать предобученный энкодер, стандартные блоки cross-attention и существующие алгоритмы выборки.

В рассматриваемой работе сравниваются два текстовых условия:

$$c_{\text{raw}} = F, \qquad c_{\text{struct}} = \text{Serialize}(g(F)), \tag{1.8}$$

где F - исходная секция *FINDINGS*, а $g(F)$ - структурированный список клинических находок. Оба варианта передаются через один тип текстового энкодера и механизм cross-attention.

Такой эксперимент не сравнивает текст с нетекстовой разметкой. В обоих случаях используется последовательность токенов, поэтому сохраняются одинаковые ограничения максимальной длины и архитектуры энкодера. Разница состоит в количестве нерелевантных фрагментов, вариативности формулировок и явности атрибутов. Благодаря этому различие результатов может быть связано именно с формой представления клинической информации, а не с изменением архитектуры диффузионной модели.

В медицинском домене проблема текстового представления особенно выражена из-за расхождения между естественным языком общего назначения

и языком клинических отчётов. Радиологические описания используют специализированные сокращения, устойчивые шаблоны отрицаний, термины локализации и сравнительные конструкции. Поэтому качество связи изображения и текста зависит не только от визуальной архитектуры, но и от того, насколько текстовый энкодер способен различать эти семантические отношения.

Развитие медицинских vision-language моделей подтверждает полезность совместного обучения по изображениям и отчётам. ConVIRT использует двунаправленный контрастивный критерий для обучения медицинских визуальных представлений по естественно существующим парам “изображение–текст” [36]. GLoRIA дополняет глобальное согласование локальным сопоставлением фрагментов изображения и слов отчёта [37]. BioViL отдельно подчёркивает роль корректного моделирования медицинской текстовой семантики и показывает, что специализированное языковое представление улучшает мультимодальное обучение [38]. В MedCLIP рассматривается возможность обучения на непарных медицинских изображениях и текстах с учётом семантических ложноотрицательных примеров [39].

Для настоящей работы эти исследования имеют два следствия. Во-первых, радиологический текст действительно содержит информацию, пригодную для формирования визуальных представлений, и потому является естественным условием генерации. Во-вторых, свободный отчёт не следует считать нейтральным контейнером признаков: его синтаксис, избыточность и отрицания непосредственно влияют на мультимодальное соответствие. Структурирование условия можно рассматривать как предобработку, уменьшающую вариативность текстовой поверхности при сохранении клинического содержания, которое должно быть передано в изображение.

При этом предлагаемый подход отличается от переобучения медицинского текстового энкодера. Специализированный энкодер меняет функцию $\tau_\eta(c)$ и может улучшить все текстовые представления одновременно; структурирование изменяет сам объект c до кодирования. Эти направления совместимы, но в контролируемом эксперименте их необходимо разделять. В настоящей работе текстовый энкодер сохраняется одинаковым для двух моделей, что позволяет исследовать именно вклад формы условия.

Важным ограничением текстового conditioning является то, что текстовый энкодер преобразует последовательность токенов в непрерывные векторы, но сам формат входа обычно не содержит явного указания на логическую структуру клинического высказывания. Cross-attention способен устанавливать мягкие соответствия между токенами и визуальными признаками, однако он не гарантирует правильного разрешения области действия отрицания, отношения атрибута к конкретной находке или различия между текущим и историческим состоянием. Поэтому увеличение мощности генеративной архитектуры не устраняет автоматически неоднозначность входного текста.

С этой точки зрения можно различить два подхода к повышению качества условной генерации. Первый направлен на улучшение самого мультимодального энкодера: использование большего корпуса, медицинского предобучения или более мощной языковой архитектуры. Второй направлен на преобразование входного сообщения до энкодера. В первом случае изменяется отображение τ_η, во втором - структура аргумента c. Предлагаемый в работе метод относится ко второму классу. Его преимущество для экспериментального анализа состоит в том, что текстовый энкодер можно сохранить неизменным и тем самым непосредственно сравнить два представления одной клинической информации.

Структурирование также связано с принципом композиционности условия. Если каждая находка представлена отдельным элементом с явно заданными атрибутами, условие естественно интерпретируется как композиция нескольких клинических требований. При свободном тексте границы таких требований задаются синтаксисом предложения и могут варьироваться от отчёта к отчёту. Фиксированный формат не гарантирует идеального выполнения каждого требования генератором, но уменьшает вариативность способа их предъявления модели. Это особенно важно при сравнении изображений, сгенерированных по однотипным клиническим описаниям.

Наконец, компактность условия имеет вычислительное следствие. Стоимость само-внимания текстового энкодера и cross-attention зависит от длины текстовой последовательности, хотя в Stable Diffusion она сверху ограничена фиксированным максимумом. Сокращение текста в пределах этого лимита не меняет асимптотическую сложность модели, но уменьшает долю контекста, занятую нерелевантными токенами.

1.3 Генерация медицинских изображений

Генерация медицинских изображений требует воспроизводить не только общий вид данных, но и клинически значимые признаки. Синтетическое изображение должно быть анатомически правдоподобным, соответствовать модальности и протоколу исследования, а в условной постановке - отражать заданное описание. Визуальная реалистичность сама по себе недостаточна: изображение может выглядеть как рентгенограмма, но не содержать указанную патологию либо воспроизводить её в неверной области.

В настоящей работе основное внимание уделяется рентгенограммам органов грудной клетки, для которых доступны крупные наборы пар “изображение - радиологический отчёт”. Дополнительно рассматриваются маммографические изображения как пример переноса подхода на другой тип медицинских данных.

1.3.1 Особенности медицинских данных

Высокое разрешение и малые диагностические признаки. Медицинское изображение содержит глобальную анатомию и локальные изменения. На рентгенограмме необходимо корректно воспроизводить лёгочные поля, сердечно-сосудистую тень, диафрагму, костные структуры и медицинские устройства. При этом клинически значимый признак может занимать небольшую часть кадра. Для маммографии особенно важны края образований, локальная перестройка ткани и микрокальцинаты. Снижение разрешения упрощает обучение, но может удалять именно те детали, которые определяют диагноз.

Неоднородность протоколов. Внешний вид изображения зависит от проекции, положения пациента, параметров оборудования и цифровой постобработки. Для рентгенографии различаются заднепередняя, переднезадняя и боковая проекции; портативные исследования систематически отличаются от плановых. Маммографическое исследование обычно включает несколько проекций обеих молочных желёз. Если эти факторы не заданы явно, модель обучается на смеси распределений и может воспроизводить технические особенности конкретного источника вместо клинически значимых закономерностей.

Даже при одинаковом клиническом состоянии распределения

изображений разных медицинских центров могут различаться из-за протоколов и оборудования. Поэтому низкое расстояние до обучающего распределения одного центра не гарантирует переносимости на другой центр. В эксперименте необходимо фиксировать проекции и правила предобработки, а тестовую выборку отделять от данных, использованных для обучения генератора.

Стоимость и неопределённость разметки. Подготовка медицинской разметки требует участия специалистов. Метка может включать диагноз, локализацию, степень выраженности, контур области или категорию оценки. Интерпретация слабых или неспецифических признаков может быть неоднозначной. В крупных рентгенографических наборах метки часто извлекаются автоматически из отчётов, поэтому наследуют их неполноту, неопределённость и ошибки обработки текста.

Следует различать истинное, но непосредственно не наблюдаемое состояние y и доступную метку $\widetilde{y}$, определяемую принятой процедурой разметки. Генератор обучается на $\widetilde{y}$ либо тексте отчёта и поэтому не может автоматически исправить ошибки исходной аннотации. В настоящей работе практическая полезность сформированных изображений дополнительно проверяется в downstream-задачах на независимых реальных тестовых данных.

Дисбаланс классов. Нормальные исследования и распространённые состояния представлены существенно лучше редких патологий и необычных сочетаний признаков. При усреднённой функции потерь редкие случаи вносят малый вклад в обучение, вследствие чего генератор может игнорировать их или заменять более типичными изменениями. Условная генерация позволяет целенаправленно формировать редкие классы, но не устраняет ограничение, связанное с малым числом исходных примеров.

Зависимость изображений одного пациента. Один пациент может иметь несколько исследований и проекций. Они сохраняют устойчивые анатомические особенности и не являются независимыми объектами. Для маммографии одно исследование обычно включает проекции CC и MLO для левой и правой молочных желёз. Независимая генерация этих изображений может нарушить согласованность плотности ткани и положения патологии.

Обучающая, валидационная и тестовая выборки должны разделяться на уровне пациентов и не пересекаться по их идентификаторам. Разделение

отдельных изображений может привести к утечке информации.

Конфиденциальность. После удаления текстовых идентификаторов изображение по-прежнему может содержать индивидуальные анатомические особенности. Синтетические данные потенциально упрощают обмен между организациями, однако генеративная модель может запомнить редкие или многократно представленные обучающие примеры. Поэтому необходимы отдельные проверки сходства с обучающей выборкой, принадлежности к ней и повторной идентификации пациента.

Перечисленные свойства показывают, что одной распределительной метрики для оценки синтетических медицинских изображений недостаточно. Поэтому в настоящей работе количественные метрики качества генерации дополняются downstream-оценкой на независимых реальных данных и отдельными экспериментами по анализу конфиденциальности. Такой протокол позволяет сопоставлять варианты генерации не только по близости распределений, но и по практической полезности и признакам возможного запоминания обучающих данных.

1.3.2 Существующие подходы к генерации рентгенографических и маммографических изображений

Ранние методы медицинской генерации основывались преимущественно на VAE и GAN. Они применялись для полного синтеза изображений, локального редактирования, преобразования между модальностями и увеличения обучающих выборок. Диффузионные модели получили распространение благодаря устойчивому обучению и возможности использовать текстовые, категориальные и пространственные условия.

Следует различать полный синтез и редактирование. При полном синтезе вся анатомия создаётся из шума, поэтому изображение потенциально относится к новому виртуальному пациенту. При редактировании сохраняется большая часть реального изображения, а генератор меняет только область или признак. Второй вариант проще и обеспечивает точную локализацию, но не подходит для создания полностью независимого набора и требует учитывать конфиденциальность неизменённого контекста.

Рентгенограммы органов грудной клетки. Рентгенограмма является проекционным изображением, на котором накладываются анатомические

структуры. Для генератора важно согласованно воспроизводить геометрию грудной клетки, лёгочные поля, сердце, сосуды, кости и устройства. Условные GAN использовали метки патологий, однако такие метки обычно содержат только факт наличия признака и не описывают его сторону, область или выраженность.

Латентные диффузионные модели позволяют использовать свободный радиологический текст. Модель RoentGen адаптирует text-to-image диффузию к рентгенограммам и отчётам [40]. В Cheff применяется каскадная латентная диффузия для получения изображений повышенного разрешения [41]. Эти подходы демонстрируют возможность синтеза по тексту, но сохраняют проблему клинического соответствия: модель может сформировать реалистичное изображение, не воспроизведя часть указанного описания.

Особое значение имеет набор MIMIC-CXR, содержащий рентгенографические исследования и связанные свободные отчёты [42]. Он обеспечивает большой объём обучающих пар, но отчёт относится ко всему исследованию, а не обязательно к одной проекции, и создавался для клинической коммуникации, а не для управления генератором.

В report-to-image постановке модель фактически изучает $p(x_v \mid R, v)$, где v - проекция. Если проекция не задана явно, она восстанавливается из статистики обучающего набора. Часть находок может быть лучше видима на боковом изображении, тогда как другая - на фронтальном. Поэтому несоответствие отдельных пар не всегда означает ошибку отчёта, но создаёт шум условного обучения. Ограничение можно уменьшить фильтрацией проекций или включением v в условие.

Маммографические изображения. Маммография характеризуется высоким разрешением и сложной текстурой ткани. Применяются два основных подхода: генерация полного изображения и локальное изменение области интереса. Прогрессивно обучаемые GAN позволяют сначала моделировать общую форму, затем добавлять детали [43]. Условное заполнение области сохраняет окружающий контекст и формирует поражение внутри заданной маски [44], но не создаёт полностью нового пациента.

Для полноразмерных изображений используются каскадные и patch-based схемы. В MAMBO отдельные модели формируют глобальный контекст и высокоразрешающие фрагменты [45]. При объединении фрагментов необходимо

контролировать швы и согласованность ткани. Отдельной задачей остаётся совместная генерация нескольких проекций одного исследования.

Набор VinDr-Mammo содержит полноформатные маммограммы, категории BI-RADS, оценки плотности и разметку патологических областей [46]. В связанной с настоящей работой публикации синтетические маммограммы исследуются с точки зрения диагностической полезности и конфиденциальности [4].

Таким образом, диффузионные модели расширили возможности условного управления, но не устранили доменные ограничения: необходимость сохранять мелкие признаки, учитывать протокол исследования, согласовывать несколько проекций и проверять соответствие клиническому описанию.

Переход диффузионных моделей в медицинский домен сопровождался двумя основными направлениями: обучением моделей непосредственно на медицинских изображениях и адаптацией крупных предобученных text-to-image систем. В сравнительном исследовании Müller-Franzes и соавторов латентные диффузионные модели были сопоставлены с GAN на нескольких медицинских модальностях; результаты показали конкурентоспособность и преимущества диффузионного подхода для синтеза медицинских изображений [47]. Это подтверждает, что преимущества диффузии не ограничиваются естественными фотографиями.

Для рентгенограмм грудной клетки особенно важны модели, использующие отчёт как условие. RoentGen адаптирует предобученную латентную диффузионную модель к CXR и исследует генерацию по свободным радиологическим текстовым запросам [40]. Cheff использует каскадную латентную диффузию для высокого разрешения и также рассматривает report-to-CXR генерацию [41]. Более поздний Chest-Diffusion сочетает специализированный медицинский текстовый энкодер с облегчённым трансформерным денойзером, подчёркивая одновременно важность доменного текстового представления и вычислительной эффективности [48]. Совокупность этих работ показывает, что радиологический отчёт стал самостоятельным источником conditioning для генерации CXR, а качество его кодирования является отдельной исследовательской проблемой.

Существующие подходы, однако, в основном отвечают на вопрос, способна ли модель адаптироваться к медицинскому визуальному и текстовому домену.

В настоящей работе ставится более узкий вопрос: можно ли повысить качество условной генерации, не меняя архитектуру Stable Diffusion, а преобразовав исходный радиологический текст в более компактную и формализованную форму. Такая постановка дополняет исследования по доменной адаптации: даже хорошо адаптированный генератор получает преимущество только в той мере, в какой условие однозначно описывает требуемые визуальные свойства.

Отдельно следует учитывать, что медицинская реалистичность не тождественна фотореалистичности. Рентгенограмма может иметь правдоподобную глобальную анатомию, но содержать ошибочную локализацию катетера, несогласованный размер сердца или патологию, не соответствующую тексту. Следовательно, условная медицинская генерация требует анализа как распределительного качества, так и содержательной полезности синтетических данных. Именно поэтому в третьей главе стандартные FID/KID дополняются downstream- и privacy-экспериментами.

1.3.3 Применение синтетических данных

Наиболее распространённый сценарий - расширение реальной обучающей выборки. Пусть $\mathcal{D}_{\text{real}} = \{(x_i, y_i)\}$ и $\mathcal{D}_{\text{syn}} = \{(\widetilde{x}_j, \widetilde{y}_j)\}$. Важно, что метка $\widetilde{y}_j$ обычно берётся из условия генерации, а не из независимой разметки результата. Поэтому ошибка следования условию становится ошибкой метки и способна ухудшить downstream-модель. Прикладная модель может обучаться по взвешенному критерию

$$\mathcal{L}_{\text{mix}}(\psi) = \frac{1}{N_r} \sum_{(x_i,y_i)\in\mathcal{D}_{\text{real}}} \ell\left(f_\psi(x_i), y_i\right) + \lambda_{\text{syn}} \frac{1}{N_s} \sum_{(\widetilde{x}_j,\widetilde{y}_j)\in\mathcal{D}_{\text{syn}}} \ell\left(f_\psi(\widetilde{x}_j), \widetilde{y}_j\right),$$

где N_r и N_s - размеры реальной и синтетической частей выборки, f_ψ - прикладная модель с параметрами ψ, $\ell(\cdot,\cdot)$ - её функция потерь, а λ_{syn} регулирует вклад синтетической части. Значение параметра и отношение N_s/N_r выбираются экспериментально: чрезмерная доля синтетики может привести к усвоению искусственных артефактов.

Балансировка редких классов. Условный генератор позволяет формировать дополнительные примеры определённых патологий, локализаций и сочетаний признаков. В отличие от простого дублирования редких изображений, такой подход потенциально увеличивает вариативность. Однако

фактическое разнообразие должно проверяться отдельно: большое число почти одинаковых изображений не устраняет дисбаланс содержательно.

Предварительное обучение и аугментация. Синтетический набор может использоваться для предварительного обучения с последующей адаптацией на реальных данных либо добавляться к реальной выборке. Исследования GAN-аугментации показывают, что эффект зависит от типа данных, качества генератора и отношения синтетических примеров к реальным [49, 50]. Практическая полезность определяется только по независимой реальной тестовой выборке. Улучшение распределительных метрик генератора не гарантирует улучшения классификации или сегментации.

Также необходимо отличать увеличение формального размера выборки от увеличения её эффективной информации. Если генератор создаёт близкие варианты одних и тех же шаблонов, число файлов растёт, но модель downstream-задачи не получает новых клинических сочетаний. Поэтому downstream-эксперимент должен дополняться анализом разнообразия и ближайших соседей.

Моделирование управляемых случаев. Изменяя условие, можно формировать наборы с заданной патологией, стороной и выраженностью. Такие данные полезны для проверки чувствительности диагностических моделей и анализа редких комбинаций. При этом необходимо убедиться, что генератор изменяет целевой признак, а не одновременно несколько несвязанных характеристик.

Обмен данными и конфиденциальность. Синтетические изображения могут использоваться вместо прямой передачи реальных данных, но только после проверки риска запоминания. Отсутствие явных идентификаторов и визуальная непохожесть на исходный объект не являются формальной гарантией анонимности.

Синтетические данные могут использоваться и для формирования контролируемых тестов. Например, при фиксированной общей формулировке можно изменять сторону находки или степень выраженности и проверять реакцию классификатора. Такой эксперимент полезен только при локальности вмешательства: генератор не должен одновременно менять проекцию, возрастные признаки и общую анатомию. Поэтому такие сценарии требуют парного анализа изображений и не заменяют обычную оценку на реальных

данных.

В настоящей работе синтетические рентгенограммы применяются для сравнения исходного и структурированного текстового условия, а синтетические маммограммы - для исследования дополнения реальной обучающей выборки. Итоговая оценка включает качество генерации, downstream-результаты и конфиденциальность.

Практическая ценность синтетических медицинских данных определяется тем, могут ли они дополнять реальную выборку, не искажая целевую задачу. Простое увеличение числа изображений не гарантирует улучшения классификатора: если генератор воспроизводит преимущественно типичные примеры, синтетическая выборка может увеличить объём данных без расширения покрытия редких состояний. Если же условие позволяет целенаправленно задавать редкие признаки, синтетические изображения потенциально могут использоваться как управляемый источник дополнительных примеров. Поэтому для задач аугментации особую ценность имеет не только визуальная реалистичность, но и способность генератора устойчиво реагировать на заданное клиническое описание.

Можно выделить несколько сценариев применения синтетики. Первый - расширение обучающей выборки при ограниченном числе размеченных изображений. Второй - балансировка классов, когда генерация используется преимущественно для редких состояний. Третий - формирование контролируемых тестовых наборов, в которых можно варьировать отдельные характеристики объекта при сохранении общей структуры. Четвёртый - исследование алгоритмов в условиях ограниченного доступа к реальным медицинским данным. Во всех случаях условная модель имеет преимущество перед безусловной, поскольку позволяет определять, какие свойства должны быть представлены в синтетическом наборе.

При этом синтетические данные не следует рассматривать как независимый источник новой клинической информации. Генеративная модель аппроксимирует закономерности обучающего распределения и потому наследует ограничения, дисбалансы и ошибки исходных данных. Если некоторый клинический признак практически отсутствует в train-выборке или систематически описывается неоднозначно, одного текстового запроса недостаточно для достоверного воспроизведения соответствующей визуальной

картины. Это ограничение важно при интерпретации downstream-экспериментов: улучшение качества классификатора свидетельствует о полезности полученной синтетики в конкретном протоколе, но не означает, что генеративная модель создаёт клинические знания, отсутствующие в исходных данных.

Особый интерес представляет смешанная схема. В ней реальные изображения сохраняют связь с фактическим распределением данных, а синтетические могут расширять вариативность отдельных классов. Исследование RoentGen показало возможность использования сгенерированных CXR для downstream-классификации [40]; аналогичная логика используется в настоящей работе при дополнительной проверке на маммографических данных. Такой эксперимент важен ещё и потому, что позволяет оценивать генерацию через независимую задачу, а не только через признаки, использованные самой генеративной моделью.

Наконец, управляемая генерация создаёт возможность анализа чувствительности к условию. Даже без отдельной экспертной оценки сравнение моделей на одинаковых наборах текстовых условий требует, чтобы условие имело воспроизводимую форму. Свободный текст затрудняет такой анализ: одна и та же находка может описываться несколькими способами, а различия в длине и нерелевантных фрагментах становятся дополнительными факторами. Структурированный список уменьшает число таких факторов и тем самым делает экспериментальное сравнение более контролируемым.

1.4 Радиологические описания как условие генерации

Радиологический отчёт содержит описание наблюдаемых признаков, их локализации и клинической интерпретации. Поэтому он является естественным источником условия для генерации медицинского изображения. Вместе с тем отчёт создаётся для передачи результата исследования врачу, а не для управления генеративной моделью. В нём сочетаются визуальные признаки, техническая информация, сравнение с предыдущими исследованиями и диагностические выводы.

Пусть $R^{(i)}$ - отчёт, связанный с набором изображений исследования $X^{(i)} = \{x_1^{(i)}, \ldots, x_{V_i}^{(i)}\}$. Условие для отдельной пары формируется процедурой $\mathcal{P}$, применяемой к отчёту. Выбор $\mathcal{P}$ определяет, какая часть клинической

информации будет передана модели и насколько однозначно она связана с текущим изображением.

1.4.1 Структура радиологического отчёта и секция FINDINGS

Типичный радиологический отчёт включает показания к исследованию, описание техники, сведения о сравнении, наблюдаемые признаки и итоговое заключение. Стандартизованные шаблоны и радиологические словари применяются для унификации структуры и терминологии [51, 52].

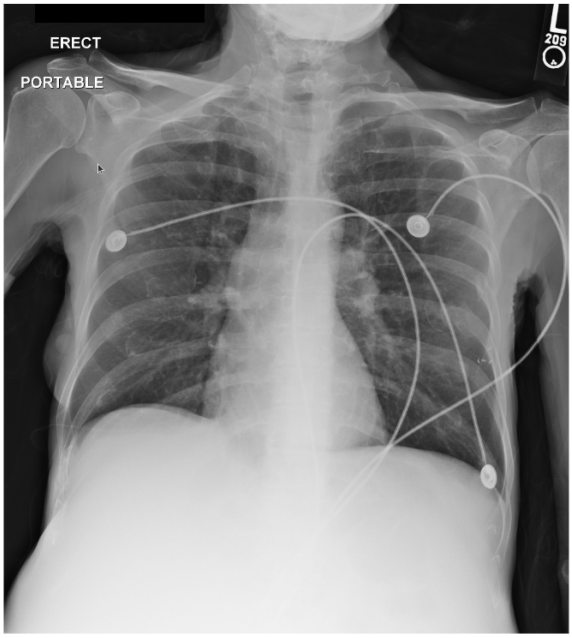
Секция *INDICATION* содержит причину назначения исследования и предполагаемый диагноз. Упомянутое заболевание может отсутствовать на изображении, поэтому эта секция не является надёжным описанием результата. *TECHNIQUE* задаёт проекцию и условия получения изображения; она полезна для контроля технического вида, но почти не описывает патологию. *COMPARISON* ссылается на предыдущие исследования и без соответствующего изображения часто не позволяет восстановить абсолютное состояние. *IMPRESSION* содержит краткий диагностический вывод, но может опускать второстепенные и нормальные признаки.

Секция *FINDINGS* наиболее непосредственно описывает визуальное содержание текущего исследования: состояние органов, патологические изменения, локализацию, выраженность и медицинские устройства.

В базовом варианте используется исходная секция *FINDINGS*, то есть условие c_{raw} из формулы (1.8). Такой выбор исключает значительную часть анамнеза и организационного текста, но сохраняет исходную формулировку врача.

Процедура извлечения должна учитывать альтернативные заголовки, регистр, переносы строк и отсутствие следующей секции. Если *FINDINGS* не выделена явно, возможны три стратегии: исключить отчёт, использовать объединённый текст либо применять дополнительный классификатор границ. Стратегия должна быть фиксирована до обучения, поскольку смешивание *FINDINGS* и *IMPRESSION* изменяет статистику условия.

В MIMIC-CXR одному отчёту может соответствовать несколько проекций [42]. Следовательно, текст описывает исследование в целом, а не является точной аннотацией каждого изображения. При подготовке пар один и тот же $F^{(i)}$ может сопоставляться фронтальной и боковой проекциям. Это




```
                                FINAL REPORT
 INDICATION:  ___ year old woman with cirrhosis.

 TECHNIQUE:  Frontal chest radiographs were obtained with the patient in the
 upright position.

 COMPARISON:  Radiographs from ___, ___ and ___.

 FINDINGS:

 The lungs are clear of focal consolidation, pleural effusion or pneumothorax.
 The heart size is normal.  The mediastinal contours are normal. Multiple
 surgical clips project over the left breast, and old left rib fractures are
 noted.

 IMPRESSION:

 No acute cardiopulmonary process.
```

Рисунок 3 — Пример пары “изображение - радиологический отчёт” из набора MIMIC-CXR

создаёт дополнительную неопределённость, которую необходимо учитывать при интерпретации результатов генерации. Пример пары “изображение - радиологический отчёт” из MIMIC-CXR показан на Рисунке 3.

Извлечение секции также является нетривиальной технической задачей: заголовки могут отсутствовать, объединяться или записываться в разных форматах. Процедура должна корректно определять начало и конец секции и одинаково обрабатывать все отчёты.

1.4.2 Недостатки неструктурированного текстового условия

Свободный текст сохраняет клиническую гибкость, но не разделяет явно тип находки, полярность, локализацию и степень выраженности. Клиническое утверждение можно представить кортежем

$$u_j = (\tau_j, p_j, \lambda_j, \rho_j, t_j)\,,$$

где τ_j - тип признака, p_j - наличие, отсутствие или неопределённость, λ_j - локализация, ρ_j - выраженность, t_j - временная характеристика. В отчёте эти компоненты выражены лексикой и синтаксисом, поэтому текстовый энкодер должен восстанавливать их неявно.

Синонимия и вариативность. Одинаковое состояние описывается

полными терминами, сокращениями, описательными фразами и локальными шаблонами учреждения. Общий языковой энкодер может по-разному представлять редкие медицинские термины. Нормализация по контролируемому словарю уменьшает эту вариативность, но требует сопоставления свободных выражений с каноническими понятиями.

Отрицания. Радиологические отчёты содержат много отрицательных утверждений, например: *No focal consolidation, pleural effusion, or pneumothorax.* Названия патологий присутствуют в тексте, но все они имеют статус *absent*. Требуется не только найти маркер отрицания, но и определить область его действия. Правиловые методы, такие как NegBio, используют синтаксические зависимости для обработки отрицаний и неопределённости [53]. Ошибка полярности приводит к условию, противоположному исходному смыслу.

С точки зрения генерации возможны две политики. Первая сохраняет отрицательные находки явно, что помогает модели не формировать указанные патологии, но увеличивает длину условия и число частых шаблонов. Вторая исключает отрицательные элементы и оставляет только наблюдаемые изменения. Она компактнее, однако отсутствие элемента становится неотличимым от отсутствия упоминания в отчёте. Выбранная политика должна быть согласована с форматом обучающих данных и одинаково применяться при генерации.

Неопределённость. Формулировки “возможно”, “не исключается” или “подозрительно на” не задают однозначного наличия признака. Возможные стратегии - сохранить статус *uncertain*, преобразовать его в положительный или отрицательный класс либо исключить находку. Политика должна быть заранее определена и одинаково применяться ко всем отчётам.

Сравнение и динамика. Выражения “без изменений”, “уменьшилось” и “больше не определяется” описывают отношение между исследованиями. Без предыдущего изображения они задают текущую картину только частично. Например, фраза о ранее описанном, но теперь не наблюдаемом признаке содержит название патологии, однако её актуальная полярность отрицательная. Для генерации текущего изображения сравнительное утверждение необходимо преобразовать в абсолютное состояние.

Нерелевантные фрагменты. Даже в *FINDINGS* встречаются сведения о качестве исследования, ограничениях интерпретации, технических

особенностях и рекомендациях. Они могут коррелировать с изображениями, но не задают требуемый визуальный признак. Если такие фрагменты сохраняются, модель может использовать побочные статистические закономерности.

Шаблонность и повторения. Стандартные отрицательные фразы встречаются значительно чаще редких патологических описаний. Повторение одного признака в нескольких предложениях также неявно увеличивает его вклад в текстовое условие. В результате частые шаблоны могут доминировать в cross-attention.

Противоречия и неполнота. После редактирования шаблона в одном отчёте могут остаться несовместимые утверждения. Обратная проблема состоит в том, что отсутствие упоминания не всегда означает отсутствие признака: врач может не описывать второстепенные изменения. Поэтому отчёт не является исчерпывающей пиксельной разметкой.

Ограничение длины. Текстовый энкодер принимает не более S токенов. Длинная секция может усекаться, а большое число нерелевантных токенов участвует в нормировке внимания. Компактное представление уменьшает риск потери важных находок и конкуренцию между клинически значимыми и шаблонными фрагментами.

Перечисленные свойства мотивируют преобразование исходной секции в структурированный список, введённое в формуле (1.8). Оно должно сохранить визуально значимую семантику, уменьшив вариативность и избыточность формы. Важно, что цель такого преобразования не состоит в получении более красивого или грамматически естественного текста. Его задача - сформировать повторяемое вычислительное представление, в котором семантически значимые различия сохраняются, а стилистические различия минимизируются.

С точки зрения обработки текста секция *FINDINGS* представляет собой последовательность клинических утверждений, в которых одно предложение может одновременно содержать несколько наблюдений и несколько областей отрицания. Например, конструкция “no focal consolidation, pleural effusion or pneumothorax” содержит три отрицаемых признака, тогда как последующее предложение может описывать положительную находку и её локализацию. Простой поиск ключевых слов не различает такие случаи и может включить в условие термин, который в отчёте фактически отрицается. Именно поэтому

системы автоматического извлечения клинических меток отдельно моделируют отрицания и неопределённость [53, 54].

Другой источник неоднозначности - временной и сравнительный контекст. Фразы “previously seen”, “resolved”, “unchanged” и “new” описывают отношение текущего изображения к предыдущим исследованиям. Для врача эта информация важна, однако генеративная модель синтезирует одно изображение текущего состояния. Следовательно, метод формирования условия должен отделить сведения о фактически видимом признаке от текста, описывающего его историю. Аналогичная проблема возникает с техническими предложениями: тип проекции, качество вдоха или наличие сравнительного исследования могут находиться рядом с описанием патологии, но не являются равноправными клиническими находками.

Наконец, свободный отчёт имеет переменную длину. Для текстового энкодера с фиксированной максимальной последовательностью это означает, что длинные описания подвергаются обрезке. Поскольку токенизация выполняется на уровне подслов, число токенов не совпадает с числом слов и особенно возрастает для редких терминов и нестандартных сокращений. Если содержательная находка расположена в конце длинной секции, она может не попасть в контекст U-Net. Структурирование частично решает эту проблему за счёт удаления повторных, отрицательных, исторических и служебных фрагментов и сохранения только компактных нормализованных элементов. Поэтому распределение длины исходных и структурированных условий является самостоятельной характеристикой метода и анализируется в экспериментальной части.

При формализации важно не потерять неопределённость, присущую медицинскому описанию. Утверждение “may represent atelectasis” не эквивалентно ни достоверному наличию ателектаза, ни его отсутствию. Поэтому бинарная схема “есть/нет” недостаточна для точного переноса семантики отчёта. В предлагаемом представлении положительно сформулированная неопределённость сохраняется отдельным статусом, а отрицательные признаки используются на этапе анализа, но не усиливают конечное условие генерации. Такой подход уменьшает вероятность того, что модель будет ориентироваться на саму лексему патологии независимо от области действия отрицания.

В совокупности эти особенности показывают, что переход от радиологического отчёта к условию генерации является самостоятельной алгоритмической задачей. Он включает не только выделение секции, но и нормализацию, анализ сущностей и атрибутов, обработку отрицаний, фильтрацию нерелевантного контекста и формирование компактного представления. Во второй главе эти операции формализуются как последовательность операторов метода разложения текстового условия.

1.4.3 Подходы к структурированию медицинского текста

Структурирование может выполняться при создании отчёта либо после него. В первом случае применяются стандартизованные шаблоны и контролируемые словари. Такой подход обеспечивает единообразие, но требует изменения клинического процесса и не решает обработку существующих архивов.

Автоматическое преобразование существующего текста обычно включает последовательность этапов

$$F \xrightarrow{\text{normalize}} F' \xrightarrow{\text{entities}} E \xrightarrow{\text{attributes}} E' \xrightarrow{\text{resolve}} A \xrightarrow{\text{serialize}} c_{\text{struct}}.$$

Здесь F - исходная текстовая секция *FINDINGS*, F' - нормализованный текст, E - набор извлечённых клинических сущностей, E' - сущности с определёнными атрибутами, A - итоговый список клинических находок, а c_{struct} - его сериализованное текстовое представление. Нормализация приводит регистр, пунктуацию и термины к единому виду. Извлечение сущностей выделяет наблюдения и анатомические области. Затем определяются отрицания, неопределённость, локализация и выраженность. На заключительном этапе объединяются дубликаты, разрешаются конфликты и формируется однозначный линейный формат.

Для автоматического преобразования свободного текста используются правиловые и нейросетевые методы. Правиловые системы сопоставляют термины со словарями, определяют отрицания и формируют фиксированный набор меток. В CheXpert правиловый инструмент присваивает наблюдениям положительный, отрицательный, неопределённый статус или отсутствие упоминания [55]. Преимущество метода - интерпретируемость; ограничение -

зависимость от полноты правил и фиксированного списка классов.

Нейросетевые классификаторы отчётов, например CheXbert, используют контекстные языковые представления и лучше учитывают вариативность формулировок [54]. Однако выход также ограничен заранее заданными классами, а локализация и выраженность обычно теряются.

Более детальное представление - список клинических находок:

$$A = \{a_1, \ldots, a_m\}, \qquad a_j = (\tau_j, p_j, \lambda_j, \rho_j)\,. \tag{1.9}$$

Число элементов зависит от конкретного отчёта. В выражении (1.9) τ_j обозначает тип клинической находки, p_j - её полярность или статус наличия, λ_j - анатомическую локализацию или латеральность, а ρ_j - степень выраженности. Таким образом, каждый элемент явно связывает тип признака с его статусом, анатомической областью и выраженностью. Список можно сериализовать в текст фиксированного формата и передать стандартному текстовому энкодеру без изменения архитектуры генеративной модели.

Порядок элементов также следует фиксировать. Возможен порядок появления в исходном тексте либо канонический порядок по анатомическим областям и типам находок. Первый лучше сохраняет связь с отчётом, второй уменьшает вариативность сериализации. Для сравнимости экспериментов используется один порядок как при обучении, так и при генерации. Разделители полей и элементов должны быть однозначными и не совпадать с обычными терминами.

Наиболее выразительное представление - граф клинических сущностей и отношений. Вершины обозначают наблюдения и анатомические структуры, а рёбра - локализацию и другие связи. RadGraph предоставляет такую схему для радиологических отчётов [56]. Граф лучше сохраняет сложные отношения, например общую локализацию нескольких признаков или связь устройства с его положением. Однако он требует более трудного извлечения и отдельного способа кодирования для диффузионной модели. Простая линеаризация графа может снова ввести зависимость от порядка, а специализированный графовый энкодер изменит архитектуру и затруднит изолированное сравнение формы условия.

Для настоящей работы выбран список находок как компромисс между выразительностью и вычислительной простотой. Он сохраняет больше

информации, чем фиксированный вектор, но может быть линеаризован и передан через тот же текстовый канал, что и исходная секция *FINDINGS*. В отличие от свободного текста порядок полей и разделители фиксируются заранее. Это упрощает интерпретацию токенов и уменьшает влияние авторского стиля.

Конкретный формат элемента списка, правила сериализации и примеры преобразования секции *FINDINGS* рассматриваются во второй главе; пример итогового преобразования приведён в Таблице 7.

Такое представление не является пересказом отчёта естественным языком. Его назначение - сформировать стабильный интерфейс между клиническим текстом и генеративной моделью. Поэтому грамматическая естественность уступает по приоритету однозначности полей и сохранению семантики.

Содержательная разница двух представлений сведена в таблицу 3.

Таблица 3 — Сравнение неструктурированного и структурированного текстового условия

Характеристика	**Свободный текст**	**Структурированное условие**
Наличие нерелевантных фрагментов	Возможно	Минимизируется при формировании списка
Интерпретируемость	Ограниченная	Более высокая за счёт фиксированной структуры
Учёт локализации и выраженности	Неявный	Явный в отдельных полях находки
Управляемость генерации	Снижается из-за вариативности и неоднозначности	Повышается за счёт однозначного задания признаков
Пригодность для формального алгоритмического анализа	Ограниченная	Высокая

Это позволяет сравнивать два варианта условий из формулы (1.8) при неизменной архитектуре модели. Структурирование не гарантирует безошибочность: возможны пропуски и ложные сущности, ошибки полярности, локализации и объединения. Поэтому результат должен быть интерпретируемым, детерминированным и пригодным для выборочной проверки. При необходимости связь элемента списка с исходным фрагментом текста может использоваться как вспомогательная диагностическая

информация, однако она не является частью условия генерации.

Важным ограничением является потеря информации при чрезмерном упрощении. Если список содержит только положительные классы, модель перестаёт получать явные сведения об отсутствии критических патологий. Если удаляются все сравнительные конструкции, может быть потеряно актуальное состояние, скрытое в выражении динамики. Поэтому структурирование должно опираться не на механическое сокращение текста, а на преобразование клинических утверждений в абсолютную и однозначную форму.

Качество преобразования зависит от корректности извлечения сущностей, определения полярности и локализации, нормализации терминов, объединения дубликатов и разрешения противоречий. Эти этапы формализуются и рассматриваются во второй главе как части разработанного алгоритма формирования списка клинических находок.

В разрабатываемом далее методе отрицательные и неопределённые утверждения учитываются при семантическом разборе и определении полярности, однако в основное условие генерации включаются подтверждённые визуально значимые находки. Если таких находок не выделено, используется специальное условие *No findings*. Эта политика далее применяется одинаково при подготовке обучающей, валидационной и тестовой выборок.

Структурирование радиологического текста связано с более общей задачей построения представлений, сохраняющих клинические отношения при уменьшении языковой вариативности. С одной стороны, онтологии и словари, такие как RadLex, задают контролируемый словарь медицинских терминов; с другой стороны, системы извлечения сущностей и отношений позволяют преобразовывать свободный отчёт в машиночитаемую структуру. RadGraph представляет сущности и отношения между наблюдениями и анатомическими областями в виде графа [56], а методы извлечения меток, такие как NegBio и CheXbert, показывают практическую важность явной обработки отрицаний и неопределённости [53, 54].

Для генеративного conditioning полный граф не всегда необходим. Его преимущества - явные отношения и возможность сложного логического анализа - сопровождаются необходимостью дополнительного кодирования графовой структуры и согласования её с архитектурой генератора. Плоский набор бинарных меток, напротив, прост, но теряет локализацию, выраженность и

переменное число наблюдений. Список находок с фиксированными полями представляет компромисс: он сохраняет ключевые атрибуты, оставаясь обычной текстовой последовательностью и потому не требуя изменения интерфейса Stable Diffusion.

Особое значение имеет обработка отрицаний. В радиологии фраза “no pleural effusion” является содержательно значимой для клинического отчёта, но для генеративного условия её роль отличается от положительной находки. Если текстовый энкодер недостаточно устойчиво моделирует область действия отрицания, лексема, обозначающая патологию, всё равно присутствует во входной последовательности. Аналогичная проблема возникает с историческими и сравнительными утверждениями: упоминание ранее существовавшего изменения не означает его наличия на текущем изображении. Поэтому формирование условия должно отделять семантический разбор отчёта от конечного набора признаков, которые действительно должны влиять на генерацию.

Нормализация также уменьшает число поверхностных вариантов одного понятия. Для общей языковой модели выражения могут различаться на уровне токенов, хотя для задачи генерации они задают одну и ту же визуальную сущность. Приведение синонимов, локализации и выраженности к согласованной форме уменьшает эту избыточность и делает условие пригодным для формального алгоритмического анализа. Эти требования непосредственно переходят в алгоритм второй главы.

1.5 Методы оценки синтетических медицинских изображений

Оценка генеративной модели не может основываться на одной метрике, поскольку распределительная близость синтетических и реальных данных, практическая полезность изображений и риск раскрытия информации об обучающей выборке характеризуют разные свойства модели. В соответствии с задачами настоящей работы экспериментальная оценка строится по трём взаимодополняющим направлениям: количественные метрики качества генерации, downstream-оценка на независимых реальных данных и анализ конфиденциальности. Дополнительно проверяется перенос принципа структурированного задания условия на маммографические изображения.

Наиболее прямым способом клинической проверки синтетических

медицинских изображений является экспертная оценка врачами соответствующего профиля. В идеальном протоколе несколько независимых врачей-рентгенологов должны в слепом режиме оценивать анатомическую правдоподобность изображения, наличие артефактов, корректность изображённых патологических изменений и соответствие синтетического изображения заданному клиническому условию. Подобные reader study применялись, например, при оценке генерации CXR-изображений в работах RoentGen и исследованиях с визуальным тестом Тьюринга [40, 57]. Такая оценка наиболее близка к реальному клиническому критерию качества, однако плохо масштабируется: требуется время врачей узкой специализации, стоимость разметки высока, а для получения устойчивых выводов желательно привлекать нескольких специалистов и достаточно большой набор изображений.

В разделе 1.3.1 уже отмечены ограниченная доступность медицинских экспертов и высокая стоимость экспертной разметки. Поэтому в настоящей работе масштабная врачебная оценка не используется как основной экспериментальный инструмент. Вместо неё применяются более доступные и воспроизводимые методы: распределительные метрики качества, downstream-оценка на независимых реальных данных и специальные проверки конфиденциальности. Эти методы дают количественные прокси-оценки, которые можно вычислять одинаковым способом для большого числа изображений.

1.5.1 Метрики качества генерации

Для прямого сравнения базового и предлагаемого вариантов используются MSE, FID и KID. Все метрики должны вычисляться при одинаковой предобработке, одинаковом размере сравниваемых выборок и одном и том же признаковом экстракторе.

Среднеквадратичная ошибка определяется как

$$\mathrm{MSE}(x, \widetilde{x}) = \frac{1}{HWC} \sum_{h,w,k} \left(x_{hwk} - \widetilde{x}_{hwk}\right)^2 .$$

Здесь H, W и C - высота, ширина и число каналов изображения; x_{hwk} и $\widetilde{x}_{hwk}$ - значения соответствующего элемента реального и синтетического изображения,

а суммирование выполняется по всем $h = 1, \ldots, H$, $w = 1, \ldots, W$ и $k = 1, \ldots, C$. Меньшее значение соответствует меньшему попиксельному различию. Для стохастической генерации MSE не является самостоятельной мерой качества, поскольку одному условию может соответствовать несколько допустимых изображений, поэтому она рассматривается совместно с распределительными метриками.

Fréchet Inception Distance (FID) сравнивает распределения признаков реальных и синтетических изображений [58]. Если признаки аппроксимируются нормальными распределениями с параметрами (μ_r, Σ_r) и (μ_s, Σ_s), то

$$\mathrm{FID} = \|\mu_r - \mu_s\|_2^2 + \mathrm{tr}\left(\Sigma_r + \Sigma_s - 2\left(\Sigma_r^{1/2}\Sigma_s\Sigma_r^{1/2}\right)^{1/2}\right).$$

Здесь μ_r и μ_s - средние векторы признаков реальных и синтетических изображений, Σ_r и Σ_s - соответствующие ковариационные матрицы, $\mathrm{tr}(\cdot)$ - след матрицы, а матричный квадратный корень задаётся для положительно полуопределённых ковариационных матриц. Меньшее значение FID соответствует большей близости распределений. Значение метрики зависит от размера выборки и признакового экстрактора, поэтому сопоставимы только результаты, полученные по одному протоколу.

Kernel Inception Distance (KID) основана на максимальном среднем расхождении между признаками реальных и синтетических объектов [59]. Для выборок признаков $Z = \{z_i\}_{i=1}^{N_r}$ и $W = \{w_j\}_{j=1}^{N_s}$ несмещённая оценка имеет вид

$$\widehat{\mathrm{KID}} = \frac{1}{N_r(N_r-1)}\sum_{i\neq j} k(z_i, z_j) + \frac{1}{N_s(N_s-1)}\sum_{i\neq j} k(w_i, w_j) - \frac{2}{N_r N_s}\sum_{i,j} k(z_i, w_j).$$

Здесь z_i и w_j - признаковые представления реального и синтетического изображения соответственно, N_r и N_s - размеры сравниваемых выборок, а $k(\cdot,\cdot)$ - положительно определённое ядро, используемое в оценке максимального среднего расхождения. Как и для FID, меньшее значение соответствует большей близости распределений. В экспериментальной главе MSE, FID и KID используются для прямого сравнения модели, обученной по исходной секции *FINDINGS*, и модели со структурированным списком клинических находок.

Структурное сходство SSIM применяется далее в экспериментах по анализу близости изображений. Индекс учитывает локальные яркость, контраст и структуру [60]; большее значение соответствует большему структурному сходству. В задачах конфиденциальности чрезмерно высокий SSIM с конкретным обучающим изображением, напротив, может указывать на нежелательную близость.

1.5.2 Downstream-оценка синтетических данных

Downstream-оценка проверяет, содержат ли синтетические изображения признаки, пригодные для обучения прикладной модели, которая затем тестируется только на независимых реальных данных. Для рентгенограмм органов грудной клетки сравниваются классификаторы, обученные на синтетических наборах, сформированных по двум вариантам условия: исходной секции *FINDINGS* и списку клинических находок. Оценка выполняется на реальной выборке CheXpert [55].

В многометочной постановке используются Macro ROC AUC, Macro PR AUC и Macro F1. ROC AUC характеризует способность ранжировать положительные и отрицательные примеры [61], а PR AUC особенно информативна при выраженном дисбалансе классов [62]. F1 зависит от выбранного порога и отражает совместный баланс precision и recall. Макроусреднение вычисляется как

$$M_{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} M_k,$$

где K - число оцениваемых клинических признаков, а M_k - значение выбранной метрики (ROC AUC, PR AUC или F1) для k-го признака.

Для дополнительного эксперимента на VinDr-Mammo [46] используется другая постановка: синтетические изображения добавляются к реальной обучающей выборке в нескольких объёмах, а классификаторы сравниваются на одной независимой реальной тестовой выборке. Такая схема позволяет проверить, способны ли синтетические данные дополнять реальные, а также оценить переносимость принципа структурированного задания условия на другой тип медицинских изображений.

1.5.3 Оценка конфиденциальности

Синтетическое происхождение изображения само по себе не гарантирует отсутствия информации об обучающих данных. Поэтому в работе используются три экспериментальные проверки: поиск ближайших обучающих изображений, оценка возможности определения принадлежности реального изображения к обучающей выборке и повторная идентификация пациента.

Поиск ближайших обучающих изображений. Для каждого синтетического изображения определяется ближайший объект обучающей выборки в признаковом пространстве:

$$d_{\text{syn}\to\text{train}}(\widetilde{x}) = \min_{x\in\mathcal{D}_{\text{train}}} d(\widetilde{x}, x).$$

Здесь $\widetilde{x}$ - синтетическое изображение, $\mathcal{D}_{\text{train}}$ - обучающая выборка, а $d(\cdot,\cdot)$ - выбранная мера расстояния в пиксельном или признаковом пространстве. Контрольным является аналогичное расстояние $d_{\text{test}\to\text{train}}$ для независимых реальных тестовых изображений. В качестве показателей используются косинусное расстояние между признаками и SSIM. Если синтетические изображения не оказываются систематически ближе к train-выборке, чем независимые реальные изображения, это не выявляет признаков прямого воспроизведения обучающих примеров в рамках выбранного признакового представления.

Определение принадлежности к обучающей выборке. Для каждого реального train- или test-изображения находится ближайшее синтетическое изображение. Расстояние до него используется как оценка для классификации объектов на members и non-members. Подобная постановка относится к классу membership inference [63, 64]. Качество атаки характеризуется ROC AUC: значение, близкое к 0.5, означает, что выбранная мера близости практически не различает обучающие и независимые тестовые изображения. В работе рассматриваются косинусное расстояние, SSIM и MSE.

Повторная идентификация пациента. Синтетическое изображение используется как запрос к галерее реальных рентгенограмм. Проверяется, встречаются ли изображения связанного пациента среди первых результатов поиска; применяются Top-1, Top-5 и mAP. Обязательным контролем является поиск real query $\to$ real gallery: он показывает, что выбранное признаковое

пространство действительно содержит устойчивые индивидуальные признаки.

Результаты таких проверок следует интерпретировать как отсутствие или наличие выявленного сигнала в конкретном протоколе, а не как формальную гарантию анонимности. Обычное обучение диффузионной модели не обеспечивает дифференциальную приватность, а возможность извлечения отдельных обучающих примеров из генеративных моделей показана в литературе [65]. Поэтому несколько взаимодополняющих проверок используются совместно.

Сводная структура экспериментальной оценки, непосредственно соответствующая экспериментам третьей главы, приведена в Таблице 4.

Таблица 4 — Методы экспериментальной оценки, используемые в работе

Эксперимент	**Сравнение**	**Основные показатели**
Качество генерации	*FINDINGS* и список находок на общем тестовом наборе	MSE, FID, KID
Downstream-оценка CXR	Классификаторы, обученные на двух синтетических наборах, на общей реальной тестовой выборке	Macro ROC AUC, Macro PR AUC, Macro F1
Поиск ближайших train-изображений	Synthetic→train и test→train	Косинусное расстояние, SSIM
Membership inference	Train→synthetic и test→synthetic	Косинусное расстояние, SSIM, MSE, ROC AUC
Повторная идентификация	Synthetic→real gallery и real-контроль	Top-1, Top-5, mAP
Перенос на маммографию	Real и смешанные real+synthetic обучающие выборки	ROC AUC, Accuracy, Precision, Recall, Specificity

Таким образом, набор методов оценки соответствует задачам работы: прямое сравнение двух вариантов текстового условия дополняется проверкой практической полезности синтетических данных, анализом сходства с обучающей выборкой, membership inference, повторной идентификацией и экспериментом на маммографических данных.

Одномерная распределительная метрика не позволяет полностью разделить качество отдельных синтетических объектов и разнообразие всей выборки. Поэтому в литературе предложены оценки precision и recall для генеративных распределений: precision интерпретируется как мера правдоподобия синтетических объектов относительно реального многообразия, а recall - как степень покрытия реального распределения [66]. Дальнейшее

развитие этого подхода привело к метрикам density и coverage, направленным на более устойчивое разделение fidelity и diversity [67]. Эти методы полезны для диагностики режима генератора, однако, как и FID, зависят от выбранного признакового пространства.

Для медицинских изображений выбор feature encoder становится особенно важным. Признаки Inception, обученной на естественных изображениях, хорошо стандартизируют сравнение с общей литературой, но не гарантируют высокой чувствительности к небольшим клиническим изменениям. Современные исследования оценки медицинской генерации показывают, что распространённые показатели качества могут слабо реагировать на локальные морфологические ошибки и не всегда согласуются с пригодностью данных для downstream-задачи [68]. Поэтому в настоящей работе FID и KID трактуются как показатели распределительной близости по выбранному протоколу, а не как самостоятельное доказательство клинической корректности.

Это обстоятельство определяет многоуровневую схему проверки. Стандартные метрики отвечают на вопрос о глобальном качестве распределения. Downstream- эксперимент проверяет наличие в синтетических данных признаков, которые может использовать независимый классификатор. Анализ ближайших соседей, membership inference и повторная идентификация оценивают другую сторону генерации - возможное сохранение специфической информации об обучающих объектах. Ни один из этих экспериментов по отдельности не является исчерпывающим, но совместно они характеризуют качество, полезность и риски использования синтетической выборки.

Таким образом, для медицинской генерации принципиальна не максимизация одной численной метрики, а согласованность результатов нескольких независимых проверок. Такой подход соответствует цели диссертации: оценить влияние структурированного условия не только на общие показатели генерации, но и на практическую полезность синтетических данных и отсутствие выявляемых признаков нежелательного воспроизведения обучающей выборки.

1.6 Требования к разрабатываемому методу

Проведённый обзор показывает, что основной проблемой является не отсутствие механизма передачи текста в латентную диффузионную модель, а форма самого условия. Секция *FINDINGS* содержит клинически значимые сведения, но представлена свободным текстом, в котором смешиваются типы находок, отрицания, локализация, выраженность, сравнительные и нерелевантные фрагменты. Поэтому разрабатываемый метод должен преобразовывать её в компактное и однозначное структурированное условие, сохраняя сведения, непосредственно связанные с изображением.

Пусть R - полный радиологический отчёт, а F - выделенная секция *FINDINGS*. Требуется преобразовать F в список клинических находок, определённый формулой (1.9), и сформировать два варианта условия согласно формуле (1.8). К методу предъявляются следующие требования.

1. Выделение клинически значимого содержания. Метод должен выделять тип находки и связанные с ней атрибуты, включая полярность, анатомическую локализацию, латеральность и выраженность, если они присутствуют в исходном описании. Нерелевантные технические и организационные фрагменты не должны усиливать условие генерации. Отрицания, неопределённость и сравнительные конструкции должны обрабатываться по единой детерминированной политике.

2. Нормализация и однозначность структурированного представления. Синонимы, сокращения и варианты написания должны приводиться к согласованной форме. Повторные находки объединяются, а противоречащие утверждения обрабатываются по фиксированному правилу. Сериализация списка должна быть детерминированной: одинаковое структурированное представление должно приводить к одинаковой последовательности токенов при повторной подготовке данных.

3. Совместимость с латентной диффузионной моделью. Структурированное условие должно использовать тот же текстовый энкодер и тот же механизм cross-attention, что и исходная секция *FINDINGS*. Архитектура VAE, U-Net, расписание шума и критерий обучения не изменяются. Тем самым в контролируемом эксперименте основным изменяемым фактором остаётся форма текстового условия.

4. Вычислительная и пространственная эффективность.

Формирование списка является предварительным этапом и не должно изменять асимптотический порядок основных вычислительных затрат генеративной модели. Для алгоритмов разложения текста необходимо получить аналитические оценки вычислительной и пространственной сложности, а также показать, что замена *FINDINGS* на список находок не изменяет асимптотическую сложность обучения и генерации Stable Diffusion при одинаковом ограничении длины текстового входа.

5. Контролируемое сравнение качества генерации. Базовый и предлагаемый варианты должны обучаться на одинаковых изображениях и разбиениях данных при одинаковой архитектуре и общей конфигурации обучения и генерации. Сравнение выполняется по MSE, FID и KID на одном тестовом наборе. Это требование соответствует основному эксперименту по оценке влияния структурированного условия.

6. Проверка практической полезности синтетических данных. Синтетические изображения, полученные при двух вариантах условия, должны сравниваться в downstream-задаче классификации на независимых реальных данных. Дополнительно должна быть проверена применимость структурированного условия к другому типу медицинских изображений и возможность использования полученной синтетики для расширения реальной обучающей выборки.

7. Анализ конфиденциальности. Экспериментальная проверка должна включать поиск ближайших обучающих изображений для синтетических объектов, оценку возможности различить train- и test-изображения по близости к синтетическим данным и оценку риска повторной идентификации пациента. Выводы ограничиваются выбранными признаковыми представлениями и сценариями атак и не формулируются как доказательство абсолютной анонимности.

1.7 Выводы к первой главе

В первой главе рассмотрены теоретические и прикладные предпосылки разработки метода условной генерации изображений со структурированным заданием условия на примере медицинских данных. По результатам обзора можно сформулировать следующие выводы.

1. Диффузионные модели обеспечивают устойчивый денойзинговый

критерий обучения и допускают гибкое условное управление. Латентная диффузия уменьшает вычислительную стоимость за счёт переноса основного процесса в сжатое пространство, сохраняя возможность text-to-image генерации через текстовый энкодер и cross-attention.

2. В условной диффузионной модели качество результата определяется не только архитектурой денойзера и объёмом данных, но и формой представления условия. Текст является гибким интерфейсом, однако свободная форма затрудняет явное выделение локализации, выраженности, отрицаний и клинической неопределённости.
3. Радиологические отчёты являются естественным источником текстового условия для медицинских изображений. Вместе с тем секция *FINDINGS* содержит синонимы, отрицательные и сравнительные конструкции, исторические упоминания и другие элементы, которые могут быть полезны врачу, но не должны одинаково влиять на генерацию текущего изображения.
4. Существующие методы медицинского NLP и vision-language обучения подтверждают целесообразность явного моделирования медицинской текстовой семантики. Для рассматриваемой задачи рациональной формой является список находок, занимающий промежуточное положение между свободным текстом, фиксированным вектором меток и полноценным графом клинических сущностей.
5. Оценка синтетических медицинских изображений требует нескольких взаимодополняющих протоколов. FID, KID и MSE характеризуют отдельные аспекты качества, downstream-оценка - практическую полезность, а поиск ближайших объектов, membership inference и повторная идентификация - выявляемые риски сохранения информации об обучающей выборке.
6. На основании проведённого анализа сформулированы требования к методу: выделение клинически значимого содержания, обработка отрицаний и неопределённости, нормализация и компактная сериализация условия, совместимость со штатным текстовым интерфейсом Stable Diffusion, вычислительная эффективность, программная воспроизводимость и возможность контролируемого экспериментального сравнения с исходной секцией *FINDINGS*.

Глава 2. Метод условной генерации изображений со структурированным текстовым условием

В настоящей главе предлагается метод условной генерации изображений со структурированным текстовым условием. Метод включает два последовательных этапа: преобразование исходного неструктурированного текстового описания в компактное структурированное условие и использование сформированного условия в штатном механизме условной латентной диффузии. Разработка и программная реализация метода рассматриваются на примере медицинских данных: исходным текстовым описанием служит радиологический отчёт, а структурированным представлением - список клинических находок.

Основным результатом первой составляющей метода является разложение текстового условия, в рамках которого из радиологического отчёта выделяется секция *FINDINGS*, из неё извлекаются клинически значимые утверждения, выполняются обработка отрицаний, нормализация терминов и связывание признаков с анатомическими областями, после чего формируется список клинических находок. Полученное представление сохраняет сведения о типе визуально наблюдаемого признака, его статусе, локализации и степени выраженности и используется в качестве текстового условия латентной диффузионной модели.

Метод рассматривается одновременно на формальном, алгоритмическом и программном уровнях. Сначала вводятся математические представления радиологического отчёта, отдельной клинической находки и списка находок. Затем описываются входящие в метод алгоритмы выделения релевантной секции отчёта, нормализации текста, извлечения клинически значимых сущностей и их атрибутов, фильтрации нерелевантных фрагментов, объединения повторов и разрешения потенциальных противоречий. После этого приводятся аналитические оценки вычислительной и пространственной сложности, а также доказывается неизменность асимптотической сложности обучения и генерации Stable Diffusion при замене исходного текстового условия структурированным. Заключительная часть главы посвящена архитектуре и программной реализации комплекса условной генерации изображений со

структурированным текстовым условием, разработанного и исследованного на примере медицинских данных.

2.1 Постановка задачи и формальная модель текстового условия

Пусть обучающий набор содержит N радиологических исследований

$$\mathcal{D} = \left\{ \left(\mathcal{X}^{(i)}, R^{(i)} \right) \right\}_{i=1}^{N}, \qquad \mathcal{X}^{(i)} = \left\{ x_1^{(i)}, \ldots, x_{V_i}^{(i)} \right\},$$

где $R^{(i)}$ - радиологический отчёт, относящийся к i-му исследованию, $\mathcal{X}^{(i)}$ - множество связанных с ним медицинских изображений, а V_i - число изображений в исследовании. Такое представление соответствует структуре MIMIC-CXR, где одному исследованию может соответствовать несколько рентгенограмм и один общий радиологический отчёт [42].

Обозначим через $F^{(i)}$ выделенное из отчёта объективное описание визуальных находок. Основным источником $F^{(i)}$ является секция *FINDINGS*. Если явная секция *FINDINGS* в документе отсутствует, алгоритм допускает использование объективной описательной части *IMPRESSION*. Такое правило не изменяет смысл базового условия: во всех стандартных случаях им является выделенная секция *FINDINGS*, а обращение к *IMPRESSION* служит обработкой исключений, связанных с неоднородной структурой исходных отчётов.

В базовом варианте условной генерации текст $F^{(i)}$ непосредственно подаётся в текстовый энкодер:

$$c_{\text{base}}^{(i)} = F^{(i)}.$$

В предлагаемом варианте он преобразуется отображением

$$g : \mathcal{F} \to \mathcal{A}, \qquad A^{(i)} = g\left(F^{(i)} \right),$$

где $\mathcal{F}$ - множество допустимых текстовых описаний, а $\mathcal{A}$ - множество конечных списков клинических находок. После сериализации формируется структурированное условие

$$c_{\text{struct}}^{(i)} = \text{Serialize}\left(A^{(i)} \right).$$

Обе экспериментальные ветви используют один и тот же текстовый энкодер, одинаковый механизм cross-attention и одну архитектуру латентной диффузионной модели. Таким образом, изменяемым фактором является представление клинической информации в условии генерации.

Задача отображения g заключается не в постановке диагноза и не в дополнении отчёта сведениями, отсутствующими в исходном тексте. Требуется выделить визуально обоснованные утверждения и привести их к компактному формату, пригодному для алгоритмической обработки. Для сформированного списка должны выполняться следующие требования:

$$\begin{aligned} &A = g(F), \\ &\forall a \in A \quad \text{Grounded}(a, F) = 1, \\ &\text{Format}(A) = 1, \\ &\text{NoDuplicate}(A) = 1, \end{aligned}$$

где $\text{Grounded}(a, F)$ означает наличие в исходном описании основания для находки a, $\text{Format}(A)$ - соответствие фиксированной схеме полей, а $\text{NoDuplicate}(A)$ - отсутствие парафразных повторов одной и той же находки.

2.1.1 Представление радиологического отчёта

Радиологический отчёт рассматривается как последовательность токенов

$$R = (r_1, r_2, \ldots, r_L),$$

содержащая несколько функциональных секций: *INDICATION*, *TECHNIQUE*, *COMPARISON*, *FINDINGS*, *IMPRESSION* и другие фрагменты. Для задачи генерации требуется объективное описание визуального содержания текущего исследования, поэтому служебные, технические и анамнестические секции в условие не включаются.

Оператор выделения релевантного текста определим как

$$F = \text{ExtractSection}(R).$$

Алгоритм реализует следующие приоритеты. Во-первых, при наличии явного заголовка *FINDINGS* извлекается соответствующая секция независимо от

варианта написания заголовка. Во-вторых, если *FINDINGS* отсутствует, из *IMPRESSION* выделяется только объективная описательная часть, без заключительных диагностических формулировок. В-третьих, при наличии обеих секций материал из *IMPRESSION* используется только для добавления новых объективных сведений, не являющихся повтором или парафразом уже содержащихся в *FINDINGS*. Правила исключают секции *HISTORY*, *INDICATION*, *TECHNIQUE*, *COMPARISON*, *REASON FOR EXAM*, *ADDENDUM* и аналогичные служебные фрагменты.

После выделения выполняется нормализация текста

$$T = \text{Normalize}(\text{Tokenize}(F)),$$

включающая приведение пробельных символов и переносов строк к согласованной форме, устранение повторной пунктуации и формирование однострочного текстового представления. Эти операции необходимы прежде всего для того, чтобы дальнейшее извлечение сущностей выполнялось над единообразным входом и не зависело от вариантов форматирования исходного файла.

Радиологический отчёт связан с исследованием, которому может соответствовать несколько изображений. Поэтому для изображения $x_v^{(i)}$ текстовое условие $F^{(i)}$ остаётся описанием исследования в целом и может использоваться для нескольких проекций. Данная особенность исходной разметки сохраняется и после структурирования и учитывается при формировании обучающих пар.

2.1.2 Представление клинической находки

Клинической находкой в настоящей работе называется нормализованное утверждение о визуально наблюдаемом признаке, медицинском устройстве или аппаратном элементе, относящееся к текущему исследованию. Каждая находка задаётся кортежем

$$a_j = (\tau_j, p_j, \lambda_j, \rho_j),$$

где τ_j обозначает тип находки, p_j - статус наличия или отсутствия признака на этапе семантического анализа, λ_j - локализацию или латеральность, а ρ_j - степень выраженности. Такое разложение отделяет собственно клинический

признак от его атрибутов и позволяет представить содержание свободного текста в форме, пригодной для формального анализа. Близкая идея явного представления сущностей и отношений используется в RadGraph [56], однако в настоящей работе итоговое условие представлено не графом, а линейным списком находок фиксированного формата.

На этапе обработки отрицаний удобно рассматривать промежуточный статус

$$p_j^* \in \{\texttt{present}, \texttt{probable}, \texttt{absent}\}.$$

Значение `absent` используется для распознавания явно отрицательных утверждений и предотвращения их ошибочного включения как положительного условия. После фильтрации отрицательных и нормальных утверждений в итоговый сериализованный список передаются только

$$p_j \in \{\texttt{present}, \texttt{probable}\}.$$

Для степени выраженности используется фиксированное множество

$$\mathcal{Q} = \{\texttt{none}, \texttt{trace}, \texttt{small}, \texttt{mild}, \texttt{moderate}, \texttt{large}, \texttt{severe}\}.$$

Если степень не указана явно или не может быть однозначно нормализована, используется значение `none`. Для локализации также используется `none`, если она отсутствует в исходном тексте. Смысл компонентов представления приведён в Таблице 5.

Таблица 5 — Компоненты формального представления клинической находки

Поле	**Содержание**	**Допустимое значение**	**Пример**
τ	Тип наблюдаемого признака или устройства	Короткий канонический клинический термин	*atelectasis*
p	Статус находки	`present`, `probable`; отрицательный статус обрабатывается до сериализации	`present`
λ	Нормализованная анатомическая локализация	Короткая клиническая фраза или `none`	*bilateral lung bases*
ρ	Степень выраженности	`none`, `trace`, `small`, `mild`, `moderate`, `large`, `severe`	`mild`

Тип τ_j представляет нормализованное понятие, а не исходную словоформу. Оператор нормализации обозначим

$$\nu_\tau : \mathcal{T}_{\text{raw}} \to \mathcal{T},$$

где $\mathcal{T}_{\text{raw}}$ - множество поверхностных форм, а $\mathcal{T}$ - множество канонических наименований. В программной реализации правила нормализации включают, в частности, приведение вариантов *NG tube*, *nasogastric tube*, *feeding tube* и *Dobhoff tube* к *enteric tube*; вариантов *PICC*, *central venous catheter*, *central venous line* и *central catheter* - к *central line*; вариантов обозначения имплантированного порта - к *port*; вариантов обозначения кардиостимулятора - к *pacemaker*.

Для выраженности используются однозначные правила нормализации: *tiny*→`trace`, *minimal*→`small`, *slight*→`mild`, *marked*→`severe`, *mild to moderate*→`moderate`, *moderate to large*→`large`. Пограничные формулировки, например *borderline* или *top-normal*, не преобразуются автоматически в положительную патологию. Формулировки *likely*, *probable*, *may represent*, *suspicious for*, *suggestive of*, *could reflect*, *possibly* и *compatible with* соответствуют статусу `probable`, если речь идёт о положительно предполагаемой текущей находке.

2.1.3 Представление списка клинических находок

Для одного исследования структурированное условие представляется списком

$$A = \{a_1, a_2, \ldots, a_m\},$$

где m - число выделенных клинических находок. После обработки отрицаний и фильтрации в список включаются текущие положительные и вероятно положительные находки, устройства или аппаратные элементы, явно связанные с наблюдаемым изображением.

При передаче в диффузионную модель каждая находка сериализуется в строку фиксированного формата

$$\text{Ser}(a_j) = \tau_j \mid \lambda_j \mid p_j \mid \rho_j.$$

Итоговое условие формируется как

$$\text{Serialize}(A) = \text{Ser}(a_1)\ ;\ \text{Ser}(a_2)\ ;\ \ldots\ ;\ \text{Ser}(a_m), \tag{2.1}$$

то есть фактический порядок полей имеет вид

`finding | location | presence | severity`.

Если после обработки текста текущих положительных или вероятно положительных находок не осталось, используется специальное текстовое условие

$$c_{\text{struct}} = \texttt{No findings}.$$

Структурированная форма решает несколько задач одновременно. Во-первых, сокращается вариативность свободного языка за счёт канонизации терминов. Во-вторых, отдельно фиксируются локализация, неопределённость и выраженность находки. В-третьих, отрицательные, нормальные, разрешившиеся, исторические и технические утверждения не занимают текстовый контекст модели. В-четвёртых, формат допускает формальный разбор, сравнение и последующую программную обработку без необходимости восстанавливать структуру исходного предложения.

Длина условия определяется после токенизации текстовым энкодером:

$$M(A) = \left|\text{Tok}(\text{Serialize}(A))\right|.$$

Для используемой конфигурации Stable Diffusion 1.5 текстовый вход ограничен 77 токенами. Поэтому компактность списка имеет практическое значение: удаление служебных фрагментов и повторов уменьшает вероятность того, что содержательная часть условия будет удалена при обрезке. Распределения длин исходного и структурированного условия анализируются в экспериментальной главе.

2.2 Метод и алгоритмы разложения текстового условия

Преобразование радиологического отчёта в структурированное условие выполняется последовательно:

$$R \xrightarrow{\text{ExtractSection}} F \xrightarrow{\text{Tokenize,Normalize}} T \xrightarrow{\text{ExtractEntities}} E \xrightarrow{\text{BuildFindings}} A \xrightarrow{\text{Serialize}} c_{\text{struct}}.$$

Здесь из исходного отчёта R сначала выделяется релевантная текстовая секция F, после нормализации формируется последовательность T, из которой извлекаются клинические сущности E. На их основе строится список находок A, который затем сериализуется в текстовое условие c_{struct}, передаваемое в диффузионную модель.

В программной реализации эта последовательность разделена на два основных этапа обработки. На первом из полного радиологического отчёта выделяется и нормализуется секция, содержащая объективное описание изображения. На втором из полученного текста извлекаются клинические сущности и их атрибуты, выполняются фильтрация и нормализация, после чего формируется итоговый список находок. Общий алгоритм приведён в листинге 2.1.

Листинг 2.1 — Алгоритм формирования списка клинических находок

```
BuildFindingsList(R)
    F <- ExtractSection(R, "FINDINGS")
    T <- Normalize(Tokenize(F))
    E <- ExtractEntities(T)
    A <- empty set

    for each e in E do
        a <- BuildFinding(e)
        if IsValid(a) then
            A <- A union {a}
        end if
    end for

    A <- MergeDuplicates(A)
    A <- ResolveConflicts(A)

    return A
```

Назначение отдельных этапов приведено в таблице 6.

Исходным текстом для дальнейшего структурирования служит секция *FINDINGS*. При её выделении учитываются различия в регистре и оформлении заголовков радиологических отчётов. Если явная секция *FINDINGS* отсутствует, допускается использование объективной описательной части *IMPRESSION*. Если присутствуют обе секции, из *IMPRESSION* могут

Таблица 6 — Этапы метода разложения текстового условия

№	Этап	Результат	Назначение
1	Выделение секции и нормализация	F, T	Получение объективного текста *FINDINGS*; обработка исключений с *IMPRESSION*; нормализация формы текста
2	Извлечение сущностей и атрибутов	E	Выделение наблюдаемого признака, локализации, статуса и выраженности
3	Построение и фильтрация находок	A	Обработка отрицаний, исключение нерелевантных элементов, канонизация терминов
4	Объединение и разрешение противоречий	A	Устранение парафразных повторов и выбор согласованного представления
5	Сериализация	c_{struct}	Формирование строки `finding | location | presence | severity` для текстового энкодера

быть добавлены только сведения о визуальных признаках, отсутствующие в *FINDINGS*; повторное диагностическое заключение при этом не используется.

После выделения релевантной части текста удаляются служебные фрагменты, лишние пробелы и переносы строк, нормализуется пунктуация. Результат приводится к единой текстовой форме, которая используется одновременно как базовое условие генерации и как вход следующего этапа структурирования. Промежуточные результаты сохраняются в процессе пакетной обработки, поэтому подготовленные записи могут повторно использоваться при обучении и экспериментальной оценке моделей.

Из нормализованного текста выделяются клинические сущности, соответствующие наблюдаемым признакам, устройствам и аппаратным элементам. Для каждой сущности определяются тип находки, анатомическая локализация, статус и степень выраженности. Синонимичные формулировки и клинические сокращения приводятся к единому представлению. Если текст указывает на вероятное наличие признака, например с помощью формулировок *likely*, *possibly* или *may represent*, для него сохраняется статус

`probable`.

В итоговый список включаются только признаки, описанные как присутствующие или вероятно присутствующие на текущем изображении. Отрицательные и нормальные утверждения, разрешившиеся находки, сведения только об анамнезе, технические замечания и сравнительные фрагменты исключаются. Например, конструкции *no pleural effusion*, *without pneumothorax*, *not seen* и *resolved edema* не формируют элементы итогового условия. В то же время хронический или послеоперационный признак сохраняется, если в отчёте явно указано, что он наблюдается на текущем изображении, например с помощью выражений *again seen*, *redemonstrated*, *persistent* или *still present*.

Если локализация или степень выраженности признака в исходном тексте не указаны, соответствующее поле принимает значение `none`. Это позволяет сохранить фиксированную структуру записи и не вводить информацию, которой нет в исходном отчёте. Для диагностических интерпретаций используется тот же принцип: в структурированное условие включаются прежде всего непосредственно наблюдаемые признаки, а диагностическая интерпретация сохраняется только тогда, когда она явно следует из текста и содержит самостоятельную информацию.

После формирования отдельных элементов выполняется объединение повторов. Записи, соответствующие одной и той же находке и имеющие эквивалентные атрибуты, сводятся к одному элементу. Если один признак описан несколько раз с различающимися характеристиками, при разрешении конфликта приоритет имеет описание текущего исследования и более явно сформулированная визуальная характеристика. Такой порядок позволяет устранить парафразные повторы, не преобразуя при этом нормальные или неопределённые формулировки в патологические признаки.

Сформированный список сериализуется согласно формуле (2.1). Каждая находка записывается в фиксированном формате

`finding | location | presence | severity`,

а отдельные элементы разделяются символом “;”. Результат сохраняется в поле `List_of_findings` и используется далее как c_{struct}. Примеры преобразования исходного текста приведены в таблице 7.

Первый пример показывает, что положительные признаки текущего исследования сохраняются, тогда как отрицания трахеального отклонения,

Таблица 7 — Примеры преобразования секции FINDINGS в список клинических находок

Исходная секция FINDINGS	**Структурированное условие**
Hyperexpanded lungs with flattened diaphragms are unchanged. Previously described right tracheal deviation is not seen on the current study. Unchanged bilateral apical bronchial thickening. The lungs are otherwise clear without focal consolidation, pneumothorax, or effusion. The cardiomediastinal and hilar silhouettes are normal. Stable descending tortuous aorta.	`hyperexpanded lungs | bilateral lungs | present | none;` `flattened diaphragm | bilateral diaphragm | present | none;` `bronchial thickening | bilateral apical | present | none;` `tortuous aorta | descending aorta | present | none`
PA and lateral views the chest provided. Biapical pleural parenchymal scarring noted. No focal consolidation concerning for pneumonia. No effusion or pneumothorax. No signs of congestion or edema. Cardiomediastinal silhouette is stable with an unfolded thoracic aorta and top-normal heart size. Bony structures are intact.	`pleural parenchymal scarring | bilateral apices | present | none`
Portable AP chest radiograph. The lungs are relatively well expanded without focal consolidation, pleural effusion or pneumothorax. The heart is normal in size.	`No findings`

консолидации, пневмоторакса и выпота не включаются в итоговое условие. Во втором примере сохраняется наблюдаемое рубцовое изменение, а отрицательные и нормальные характеристики исключаются. В третьем случае положительные находки отсутствуют, поэтому результатом является `No findings`.

В результате преобразования свободный текст радиологического отчёта представляется в виде списка клинических находок фиксированной структуры. Алгоритм сохраняет положительные и вероятные визуальные признаки вместе с их локализацией и выраженностью, одновременно исключая отрицания, повторы, служебные и нерелевантные фрагменты. Полученное представление используется далее как структурированное текстовое условие латентной диффузионной модели.

2.3 Аналитические оценки сложности разработанного метода

В предыдущем разделе был описан метод разложения текстового условия, преобразующий радиологический отчёт в структурированный список клинических находок. Помимо корректности формируемого представления существенным свойством разработанного метода является вычислительная применимость при обработке больших наборов радиологических отчётов. Поэтому далее рассматриваются аналитические оценки вычислительной сложности формирования списка находок, пространственной сложности его хранения, а также влияние замены исходного текстового условия структурированным представлением на сложность обучения и генерации латентной диффузионной модели.

Анализ выполняется отдельно для двух частей разработанного метода. Сначала рассматривается этап предварительной обработки радиологического отчёта

$$R \longrightarrow F \longrightarrow A \longrightarrow c_{\text{struct}},$$

где R - полный радиологический отчёт, F - выделенное и очищенное описание наблюдаемой картины, A - список клинических находок, а c_{struct} - его сериализованное представление. Затем предполагается, что условие уже сформировано, и рассматривается его использование в модели Stable Diffusion.

Для аналитической оценки операции метода рассматриваются в соответствии с их алгоритмическим назначением: выделение секции, нормализация, сопоставление, классификация, фильтрация и формирование элементов структурированного представления. Такой уровень описания позволяет оценить зависимость затрат от длины входного отчёта и размера итогового списка независимо от конкретной программной платформы выполнения семантических операторов.

2.3.1 Вычислительная сложность формирования списка находок

Рассмотрим алгоритм формирования списка клинических находок. Пусть исходный радиологический отчёт представлен последовательностью

$$R = (r_1, r_2, \ldots, r_L),$$

содержащей L токенов. Из него выделяется клинически релевантная часть

$$F = (f_1, f_2, \ldots, f_l), \qquad l \leq L.$$

Обработка включает выделение релевантной секции, нормализацию текста, извлечение клинически значимых сущностей и их атрибутов, фильтрацию нерелевантных утверждений, обработку отрицаний, устранение повторов и формирование итогового списка. В программном комплексе эти этапы выполняются последовательно, а промежуточные и итоговые результаты сохраняются для повторного использования.

Для аналитической оценки введём параметр $d \geq 1$, задающий верхнюю границу числа элементарных операций сопоставления, нормализации, классификации или фильтрации, относимых к обработке одного токена в рассматриваемой алгоритмической схеме. Пусть также C обозначает число типов клинических признаков, предусмотренных фиксированной схемой структурированного представления. В аналитической модели C рассматривается как параметр выбранной схемы и не зависит от длины конкретного радиологического отчёта. При аналитической оценке семантический преобразователь рассматривается как внешний оператор, а параметр d характеризует число операций алгоритмической схемы над элементом вход.

Лемма 1. *Пусть радиологический отчёт R содержит L токенов. Тогда выделение клинически релевантной части отчёта, в которой приоритет имеет секция FINDINGS, а секция IMPRESSION используется в предусмотренных алгоритмом резервных случаях, имеет вычислительную сложность $\mathcal{O}(L)$ в принятой алгоритмической модели.*

Доказательство. Для определения границ клинически релевантной части необходимо обработать входную последовательность токенов отчёта и установить принадлежность фрагментов к предусмотренным секциям. Каждый токен исходного отчёта требуется рассмотреть не более постоянного числа раз. Наличие резервного правила для секции *IMPRESSION* не требует дополнительного асимптотически большего обхода, поскольку она является частью того же отчёта. Следовательно,

$$T_{\text{extract}}(L) = \mathcal{O}(L).$$

Что и требовалось доказать. □

Лемма 2. *Пусть выделенная клинически релевантная часть F содержит l токенов. Если для обработки одного токена выполняется не более d элементарных операций сопоставления, нормализации, классификации или фильтрации, а число типов клинических признаков в фиксированной схеме представления ограничено величиной C, то вычислительная сложность семантической обработки F оценивается как*

$$T_{\text{semantic}}(l, d, C) = \mathcal{O}(l \cdot d + C).$$

Доказательство. При последовательной обработке l токенов на каждый из них приходится не более d операций, включающих нормализацию, выделение клинически значимых сущностей и атрибутов, определение статуса присутствия, обработку отрицаний и фильтрацию нерелевантных утверждений. Поэтому стоимость обработки токенов не превосходит $\mathcal{O}(l \cdot d)$. Учёт фиксированной схемы типов клинических признаков в общем виде добавляет $\mathcal{O}(C)$. Суммирование этих слагаемых даёт

$$T_{\text{semantic}}(l, d, C) = \mathcal{O}(l \cdot d + C).$$

Что и требовалось доказать. □

Лемма 3. *Пусть после семантической обработки получено m валидных клинических находок. Каждая находка формируется из фиксированного числа полей и включается в итоговый список не более одного раза. Предположим также, что размер сериализованного представления одной находки ограничен сверху константой a, не зависящей от m и длины исходного отчёта. Тогда формирование и сериализация итогового списка имеют вычислительную сложность $\mathcal{O}(m)$. При этом в принятой алгоритмической модели $m = \mathcal{O}(l \cdot d)$.*

Доказательство. Каждая из m итоговых находок последовательно преобразуется в запись фиксированной структуры `finding | location | presence | severity`. Поскольку размер сериализованного представления одной находки ограничен константой a, формирование её полей, служебных разделителей и добавление записи в итоговое представление требуют

$\mathcal{O}(a) = \mathcal{O}(1)$ операций. Поэтому для m находок

$$T_{\text{serialize}}(m) = \mathcal{O}(m \cdot a) = \mathcal{O}(m).$$

Каждая валидная находка возникает в результате хотя бы одной операции извлечения или формирования кандидата, учитываемой среди не более чем d операций на один из l токенов. Поэтому $m = \mathcal{O}(l \cdot d)$, и стоимость сериализации не превышает порядка семантической обработки. Что и требовалось доказать. □

Теорема 1. *Пусть R - радиологический отчёт, состоящий из L токенов. Пусть из отчёта R выделяется секция FINDINGS, содержащая l токенов, где $l \leq L$. Для каждого токена секции FINDINGS выполняется не более d элементарных операций сопоставления, нормализации, классификации или фильтрации, число типов клинических признаков, используемых в схеме представления списка находок, ограничено константой C, а размер сериализованного представления одной клинической находки ограничен константой a. Тогда вычислительная сложность алгоритма формирования списка клинических находок из одного радиологического отчёта оценивается величиной*

$$T_{\text{build}} = \mathcal{O}(L + l \cdot d + C)\,. \tag{2.2}$$

Доказательство. По лемме 1 выделение клинически релевантной части имеет сложность $\mathcal{O}(L)$. По лемме 2 её семантическая обработка имеет сложность $\mathcal{O}(l \cdot d + C)$. По лемме 3 при ограниченном константой a размере сериализованного представления одной находки формирование и сериализация m итоговых находок требуют $\mathcal{O}(m)$ операций. Следовательно,

$$\begin{aligned} T_{\text{build}} &= \mathcal{O}(L) + \mathcal{O}(l \cdot d + C) + \mathcal{O}(m) \\ &= \mathcal{O}(L + l \cdot d + C), \end{aligned}$$

поскольку по лемме 3 $m = \mathcal{O}(l \cdot d)$. Что и требовалось доказать. □

Для резервного случая, когда при отсутствии явной секции *FINDINGS* используется объективная часть *IMPRESSION*, та же оценка сохраняется: обрабатываемый текст по-прежнему является частью исходного отчёта и его длина не превосходит L.

Если параметры d, C и a фиксированы выбранной схемой алгоритма и не зависят от длины отчёта, то с учётом $l \leq L$ из формулы (2.2) следует

$$T_{\text{build}} = \mathcal{O}(L).$$

Таким образом, в принятой аналитической модели вычислительная сложность формирования списка находок линейно ограничена длиной радиологического отчёта. Для набора из N отчётов с длинами $L_1, \ldots, L_N$ соответствующая оценка имеет вид

$$T_{\text{dataset}} = \mathcal{O}\left(\sum_{i=1}^{N} L_i\right).$$

Полученная оценка относится к алгоритму формирования структурированного условия как к этапу предварительной подготовки данных. В программном комплексе результат этой подготовки сохраняется и повторно используется при обучении и генерации; поэтому формирование списка не входит во внутренний цикл оптимизации Stable Diffusion.

2.3.2 Пространственная сложность хранения списка находок

Рассмотрим объём памяти, необходимый для хранения структурированного условия. Для одного исследования итоговый список имеет вид

$$A = (a_1, a_2, \ldots, a_m),$$

где m - число клинических находок. В соответствии с формальным представлением каждая находка задаётся конечным кортежем

$$a_j = (\tau_j, p_j, \lambda_j, \rho_j).$$

Число полей кортежа фиксировано и не зависит от числа находок в отчёте. Для учёта переменной длины текстовых полей обозначим через s_j суммарный размер сериализованного представления атрибутов находки a_j в машинных словах.

Лемма 4. *Для хранения одной клинической находки a_j, суммарный размер сериализованных полей которой равен s_j, требуется $\mathcal{O}(s_j)$ памяти.*

Доказательство. Находка содержит фиксированное число полей: тип находки,

статус присутствия, локализацию и степень выраженности. Служебные разделители между полями имеют постоянный суммарный размер. Поэтому объём памяти определяется суммарной длиной самих полей и отличается от s_j не более чем на постоянное слагаемое:

$$M(a_j) = \mathcal{O}(s_j + 1) = \mathcal{O}(s_j).$$

Что и требовалось доказать. □

Лемма 5. *Пусть список* $A = (a_1, \ldots, a_m)$ *содержит* m *клинических находок, а размеры их сериализованных представлений равны* $s_1, \ldots, s_m$. *Тогда пространственная сложность хранения списка оценивается как*

$$M(A) = \mathcal{O}\left(m + \sum_{j=1}^{m} s_j\right).$$

Если существует верхняя граница $s_j \leq a$ *для всех* j, *то*

$$M(A) = \mathcal{O}(m \cdot a).$$

Доказательство. По лемме 4 для хранения j-й находки требуется $\mathcal{O}(s_j)$ памяти. Кроме содержимого полей, между соседними находками сохраняются служебные разделители, суммарный размер которых линейно зависит от числа элементов и имеет порядок $\mathcal{O}(m)$. Следовательно,

$$M(A) = \mathcal{O}\left(m + \sum_{j=1}^{m} s_j\right).$$

При $s_j \leq a$ выполняется

$$\sum_{j=1}^{m} s_j \leq ma,$$

откуда следует $M(A) = \mathcal{O}(ma)$. Что и требовалось доказать. □

Теорема 2. *Пусть список клинических находок одного исследования содержит* m *элементов. Если каждая находка кодируется конечным набором атрибутов фиксированного формата и суммарный размер кодирования одной находки не превосходит* a *машинных слов, то пространственная сложность*

хранения структурированного списка оценивается как

$$M_{\text{findings}}(m, a) = \mathcal{O}(m \cdot a). \tag{2.3}$$

Доказательство. Утверждение непосредственно следует из лемм 4 и 5: суммарная память равна порядку суммы размеров всех находок и служебных разделителей, а при $s_j \leq a$ эта сумма ограничена величиной порядка ma. Следовательно,

$$M_{\text{findings}}(m, a) = \mathcal{O}(m \cdot a).$$

Что и требовалось доказать. □

В разработанном представлении число полей клинической находки фиксировано: тип находки, признак присутствия, локализация и степень выраженности. Если длины их сериализованных представлений ограничены сверху и $a = \text{const}$, то из теоремы 2 и оценки (2.3) следует

$$M_{\text{findings}}(m) = \mathcal{O}(m).$$

Следовательно, при указанном условии объём памяти, требуемый для хранения структурированного условия, линейно ограничен количеством выделенных находок.

Итоговое условие сохраняется как сериализованная строка в CSV. Общая оценка из леммы 5 применима и к такому представлению, в том числе когда текстовые поля имеют различную длину. Частный вывод $\mathcal{O}(m)$ используется при ограниченной сверху длине одной записи. Ограничение в 77 токенов применяется позднее, при передаче текста в энкодер Stable Diffusion 1.5, и не является ограничением на размер сохранённой строки в CSV.

2.3.3 Асимптотическая сложность обучения и генерации при использовании структурированного условия

Полученные выше оценки относятся к предварительному формированию структурированного условия. Далее необходимо определить, изменяет ли замена исходной секции *FINDINGS* списком клинических находок вычислительную сложность непосредственно генеративной модели.

В настоящей работе оба варианта условия передаются в Stable Diffusion

через одинаковый текстовый интерфейс. Для базового варианта

$$c_{\text{base}} = F,$$

а для предлагаемого

$$c_{\text{struct}} = \text{Serialize}(A).$$

Обозначим длины последовательностей после токенизации как

$$s_{\text{base}} = \left|\text{Tok}(c_{\text{base}})\right|,$$
$$s_{\text{struct}} = \left|\text{Tok}(c_{\text{struct}})\right|.$$

Текстовый энкодер имеет фиксированную максимальную длину входа S, поэтому для обеих постановок

$$s_{\text{base}} \leq S, \qquad s_{\text{struct}} \leq S.$$

В использованной конфигурации Stable Diffusion 1.5 значение S равно 77 токенам.

Для анализа введём следующие обозначения:

- F_V - вычислительная сложность операций вариационного автоэнкодера, необходимых для обработки одного изображения;
- $F_E(s)$ - вычислительная сложность текстового энкодера для последовательности длины s;
- $F_U(S)$ - вычислительная сложность одного прямого прохода U-Net, включая cross-attention при условии длины не более S;
- F_B - вычислительная сложность обратного распространения ошибки при одном шаге обучения;
- T - число шагов обратного диффузионного процесса при генерации одного изображения.

Лемма 6. *Пусть два текстовых условия c_1 и c_2 обрабатываются одним и тем же текстовым энкодером, а длины их токенизированных представлений удовлетворяют $s_1 \leq S$ и $s_2 \leq S$. Если $F_E(s)$ не убывает с ростом длины входной последовательности, то вычислительная сложность кодирования каждого из условий ограничена величиной $\mathcal{O}(F_E(S))$. Следовательно, замена одного условия другим при общей верхней границе S*

не изменяет асимптотический порядок затрат текстового энкодера.

Доказательство. По определению $F_E(s)$ задаёт стоимость обработки последовательности длины s. Для $s_i \leq S$ и неубывающей F_E выполняется $F_E(s_i) \leq F_E(S)$, $i \in \{1, 2\}$. Поэтому для обоих вариантов

$$T_{\text{text}} = \mathcal{O}(F_E(S)).$$

Использование одинаковой архитектуры энкодера исключает появление нового слагаемого, зависящего от формы текстового условия. Что и требовалось доказать. □

Лемма 7. *Пусть при замене базового текстового условия структурированным сохраняются архитектуры VAE, текстового энкодера и U-Net, механизм cross-attention, scheduler и число шагов обратного диффузионного процесса T. Тогда число вызовов основных вычислительных компонентов при одном шаге обучения и при генерации одного изображения не изменяется.*

Доказательство. При одном шаге обучения в обеих постановках выполняются: обработка изображения VAE, однократное кодирование текстового условия, прямой проход U-Net, вычисление функции потерь и обратное распространение ошибки. Следовательно, структура вычислительного графа на уровне перечисленных компонентов одинакова. При генерации текстовое условие кодируется один раз, затем U-Net вызывается на каждом из T шагов обратной диффузии, после чего полученное латентное представление декодируется VAE. Поскольку указанные компоненты и T одинаковы, число их вызовов не зависит от содержательной формы условия. Что и требовалось доказать. □

Теорема 3. *Пусть в латентной диффузионной модели типа Stable Diffusion исходная секция FINDINGS заменяется структурированным текстовым представлением в виде списка клинических находок. Пусть оба варианта условия передаются через один и тот же текстовый энкодер и один и тот же механизм текстового задания условия, причём длина каждого текстового входа ограничена сверху константой S. Тогда сложность одного шага обучения имеет порядок*

$$T_{\text{train}} = \mathcal{O}\left(F_V + F_E(S) + F_U(S) + F_B\right), \tag{2.4}$$

а сложность генерации одного изображения при T шагах обратного диффузионного процесса имеет порядок

$$T_{\text{gen}} = \mathcal{O}\left(F_E(S) + T \cdot F_U(S) + F_V\right). \tag{2.5}$$

Следовательно, замена секции FINDINGS на список клинических находок не изменяет асимптотическую сложность обучения и генерации модели Stable Diffusion.

Доказательство. По лемме 6 кодирование каждого из двух текстовых условий имеет порядок $\mathcal{O}(F_E(S))$. По лемме 7 последовательность вызовов основных вычислительных компонентов при обучении и генерации одинакова для базового и структурированного вариантов.

Рассмотрим один шаг обучения. Для обучающей пары (x, c) изображение x обрабатывается VAE со стоимостью F_V, текстовое условие кодируется со стоимостью порядка $F_E(S)$, после чего выполняется прямой проход U-Net со стоимостью $F_U(S)$ и обратное распространение ошибки со стоимостью F_B. Поэтому

$$T_{\text{train}} = \mathcal{O}\left(F_V + F_E(S) + F_U(S) + F_B\right).$$

При генерации текстовое условие кодируется однократно, после чего выполняются T шагов обратного диффузионного процесса. На каждом шаге выполняется проход U-Net, поэтому суммарная стоимость денойзинга имеет порядок $\mathcal{O}(T \cdot F_U(S))$. После завершения обратного процесса латентное представление декодируется VAE. Следовательно,

$$T_{\text{gen}} = \mathcal{O}\left(F_E(S) + T \cdot F_U(S) + F_V\right).$$

Для базового и предлагаемого вариантов используются одинаковые VAE, U-Net, текстовый энкодер и алгоритм обратной диффузии. Единственным различием является содержимое текстовой последовательности, длина которой в обоих случаях ограничена одной и той же константой S. Поэтому переход от исходной секции *FINDINGS* к сериализованному списку клинических находок не добавляет нового асимптотически растущего слагаемого ни в формулу (2.4), ни в формулу (2.5). Что и требовалось доказать. □

Для наглядности результаты анализа приведены в Таблице 8.

Таблица 8 — Асимптотические оценки сложности этапов разработанного метода

Этап	**Оценка**	**При фиксированных параметрах**
Формирование списка находок одного отчёта	$\mathcal{O}(L + l\,d + C)$	$\mathcal{O}(L)$
Хранение списка из m находок	$\mathcal{O}(ma)$	$\mathcal{O}(m)$
Один шаг обучения Stable Diffusion	$\mathcal{O}(F_V + F_E(S) + F_U(S) + F_B)$	Не изменяется при замене условия
Генерация одного изображения	$\mathcal{O}(F_E(S) + TF_U(S) + F_V)$	Не изменяется при замене условия

Полученные оценки показывают, что при фиксированной схеме представления алгоритмическая сложность формирования списка линейно ограничена длиной отчёта, хранение списка требует памяти, линейной по числу находок при ограниченной длине одной записи, а замена исходного текста структурированным условием не изменяет асимптотическую сложность обучения и генерации Stable Diffusion.

При использовании classifier-free guidance число вычислений U-Net на одном шаге генерации может увеличиваться на постоянный множитель вследствие вычисления условной и безусловной оценок шума. Этот множитель одинаков для обоих вариантов текстового условия и потому также не меняет асимптотический результат теоремы 3.

2.4 Использование структурированного условия в латентной диффузионной модели

После формирования списка находок он используется в стандартном text-to-image конвейере латентной диффузионной модели. Принципиальным свойством предлагаемого подхода является отсутствие архитектурных изменений Stable Diffusion: в базовом и предлагаемом вариантах сохраняются VAE, текстовый энкодер, U-Net, механизм cross-attention и расписание диффузионного процесса. Изменяется только содержательная форма текста, поступающего на вход энкодера. Это обеспечивает прямое сравнение двух способов представления условия при неизменной архитектуре модели.

Для двух вариантов условия используются обозначения

$$c_{\text{raw}} = F, \qquad c_{\text{struct}} = \text{Serialize}(A).$$

Оба условия после токенизации преобразуются одним и тем же текстовым энкодером в последовательность контекстных представлений и далее влияют на U-Net через cross-attention, описанный в подразделе 1.2.2.

2.4.1 Ограничение длины текстового входа и обрезка токенов

Текстовый интерфейс Stable Diffusion имеет фиксированную максимальную длину последовательности, определяемую конфигурацией используемого текстового энкодера и токенизатора [24, 35]. В использованной конфигурации Stable Diffusion 1.5 максимальная длина составляет $S = 77$ токенов. Если после токенизации исходная секция *FINDINGS* содержит больше 77 токенов, хвост последовательности обрезается и часть описания фактически не участвует в формировании cross-attention контекста.

Предлагаемый метод уменьшает эту проблему не за счёт изменения лимита модели, а за счёт предварительного удаления нерелевантных фрагментов и сериализации только подтверждённых клинически значимых находок. Поэтому для одной и той же исходной секции, как правило, выполняется

$$|\,\text{Tok}(c_{\text{struct}})| <= |\,\text{Tok}(c_{\text{raw}})|,$$

что снижает число случаев, в которых условие приходится обрезать до S токенов. Это свойство имеет самостоятельное практическое значение: модель получает более компактное условие, а клинически значимые элементы реже оказываются за границей допустимого текстового входа.

Распределения длины условий - включая среднее, медиану, 95-й процентиль и долю строк, превышающих ограничение 77 токенов, - рассматриваются в Главе 3 как дополнительная характеристика преобразования текстового условия.

При сравнении двух вариантов обрезка выполняется одним и тем же токенизатором и одним и тем же правилом truncation. Следовательно, различие в количестве обрезанных токенов является следствием именно преобразования условия, а не изменения архитектуры или параметров текстового энкодера.

В Главе 3 это различие характеризуется распределениями длины двух типов условий и долей объектов, для которых требуется обрезка.

2.4.2 Подготовка обучающих пар

Пусть $R^{(i)}$ - отчёт i-го исследования, а $\mathcal{X}^{(i)} = \{x_1^{(i)}, \ldots, x_{V_i}^{(i)}\}$ - связанные с ним изображения. Из отчёта выделяется секция $F^{(i)}$, после чего для того же изображения формируются два альтернативных условия:

$$c_{\text{raw}}^{(i)} = F^{(i)},$$
$$c_{\text{struct}}^{(i)} = \text{Serialize}\left(g(F^{(i)})\right).$$

В результате строятся два набора пар “изображение - текстовое условие”, различающиеся только второй компонентой. Множество изображений и разбиение по выборкам должны совпадать, что позволяет связывать различия результатов именно с формой представления условия.

Сравнение вариантов приведено в Таблице 9.

Таблица 9 — Формирование обучающих пар для двух вариантов текстового условия

Характеристика	**Базовый вариант**	**Предлагаемый вариант**
Изображение	То же $x_v^{(i)}$	То же $x_v^{(i)}$
Источник текста	Секция *FINDINGS*	Та же секция *FINDINGS*
Условие	$F^{(i)}$	Serialize($g(F^{(i)})$)
Текстовый энкодер	Общий	Общий
Архитектура Stable Diffusion	Без изменений	Без изменений

Сформированное структурированное условие является результатом предварительной обработки и может быть подготовлено до запуска обучения. Конкретные правила предобработки изображений, характеристики выборок и гиперпараметры относятся к экспериментальному протоколу и приводятся в Главе 3.

Общая схема формирования структурированного условия и его использования при обучении Stable Diffusion приведена на Рисунке 4.

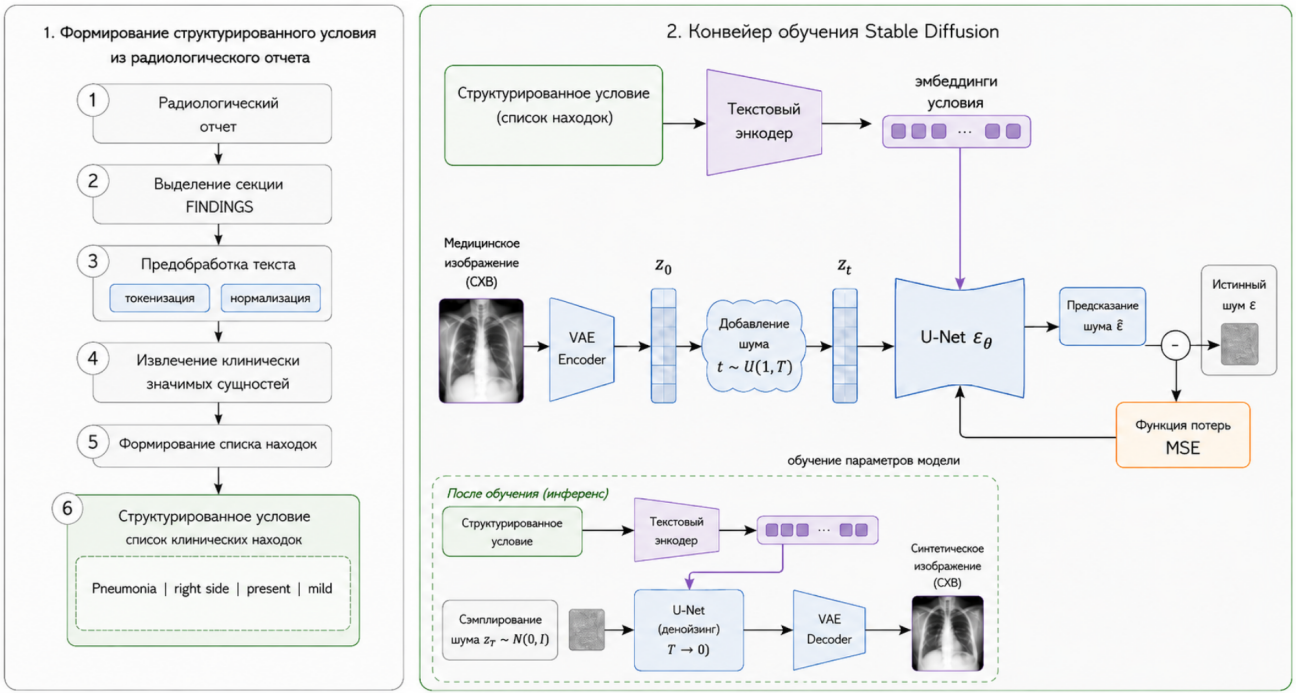


Рисунок 4 — Схема формирования и использования структурированного условия в конвейере обучения Stable Diffusion

2.4.3 Обучение модели

Обучение выполняется по стандартному критерию условной латентной диффузии, введённому в формуле (1.7). Для компактного обозначения двух постановок положим $b \in \{\texttt{raw}, \texttt{struct}\}$ и запишем

$$\mathcal{L}^{(b)}(\theta) = \mathbb{E}\left[\left\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau_\eta(c^{(b)})\right)\right\|_2^2\right].$$

Прямой диффузионный процесс, способ кодирования изображения и функция потерь не изменяются. Таким образом, предлагаемый метод не вводит дополнительный классификатор или новый член функции потерь; изменение происходит до текстового энкодера и состоит в замене свободной секции *FINDINGS* на сериализованный список клинических находок.

2.4.4 Генерация изображений

После обучения список A^* сериализуется по правилу формулы (2.1) и кодируется тем же текстовым энкодером. Обратный диффузионный процесс начинается со стандартного гауссовского шума и использует полученное текстовое представление на каждом шаге денойзинга. Если применяется classifier-free guidance, используется та же схема условного и безусловного

предсказания, которая рассмотрена в подразделе 1.2.2; меняется только строка условной ветви.

Полный путь можно кратко представить как

$$A^* \xrightarrow{\text{Serialize}} c^* \xrightarrow{\tau_\eta} Y^*, \qquad z_T \sim \mathcal{N}(0, I) \xrightarrow{\text{денойзинг при } Y^*} z_0 \xrightarrow{\mathcal{D}_\omega} \widetilde{x}.$$

При сравнении двух моделей число шагов генерации, scheduler, коэффициент guidance и прочие параметры выборки фиксируются одинаково.

2.5 Архитектура и программная реализация комплекса условной генерации

Предложенный метод реализован в виде программного комплекса, объединяющего подготовку данных, формирование текстового условия, обучение и применение латентной диффузионной модели, а также экспериментальную оценку полученных синтетических изображений. Структурирование текста выполняется до передачи условия в Stable Diffusion, поэтому архитектура самой латентной диффузионной модели не изменяется [24]. Это позволяет использовать одинаковый генеративный конвейер как для исходной секции *FINDINGS*, так и для предлагаемого структурированного условия.

Программный комплекс состоит из шести логических модулей. Первые три обеспечивают подготовку исходных данных и текстового условия, следующие два отвечают за обучение и генерацию, а последний используется для экспериментальной оценки. Назначение модулей и передаваемые между ними данные приведены в Таблице 10.

Обработка начинается с формирования записей, связывающих изображение MIMIC-CXR с соответствующим радиологическим отчётом и принадлежностью к обучающей, валидационной или тестовой выборке. Из полного текста отчёта выделяется секция *FINDINGS*, после чего для предлагаемого варианта она преобразуется в список клинических находок по алгоритму, описанному в разделе 2.2.

Промежуточные результаты текстовой обработки сохраняются в CSV-файлах. Для каждого исследования сохраняются исходный текст, результат выделения *FINDINGS*, сформированный список находок и

Таблица 10 — Модули программного комплекса и их интерфейсы

Модуль	**Назначение**	**Вход**	**Выход**
Dataset preparation	Связывание исследования, изображений, отчёта и split	метаданные MIMIC-CXR, изображения, отчёты	записи исследования
Модуль выделения FINDINGS	Выделение объективного описания из полного отчёта	текст отчёта	`model_output`, `model_error`
Модуль формирования списка находок	Формирование списка находок фиксированной структуры	`model_output`	`List_of_findings`
Diffusion training	Обучение двух вариантов при общей конфигурации	(x, c), параметры обучения	checkpoint, журнал обучения
Generation	Синтез изображений по заданному условию	checkpoint, c, параметры генерации	синтетические изображения
Evaluation	Расчёт метрик качества, downstream- и privacy-эксперименты	real/synthetic данные	таблицы метрик и промежуточные результаты

информация об ошибке обработки при её наличии. Наличие промежуточного представления позволяет повторно использовать уже обработанные записи и продолжать пакетную подготовку данных после прерывания без повторного обращения к модели для всех предыдущих исследований.

После подготовки текстового условия формируются две экспериментальные ветви. В базовой ветви условием генерации является секция *FINDINGS*; в предлагаемой ветви используется поле `List_of_findings`, содержащее сериализованный список находок. В обоих случаях на вход генеративной модели поступает пара “изображение–текст”. Используются одна архитектура Stable Diffusion 1.5 и одинаковый механизм передачи текстового условия. Тем самым различие между ветвями возникает на этапе формирования условия. Общая архитектура программного комплекса и последовательность передачи данных между его компонентами представлены на Рисунке 5.

Основная реализация комплекса выполнена на Python. Для обучения и применения диффузионной модели используются PyTorch и Diffusers, для обработки табличных данных - Pandas. Работа с медицинскими

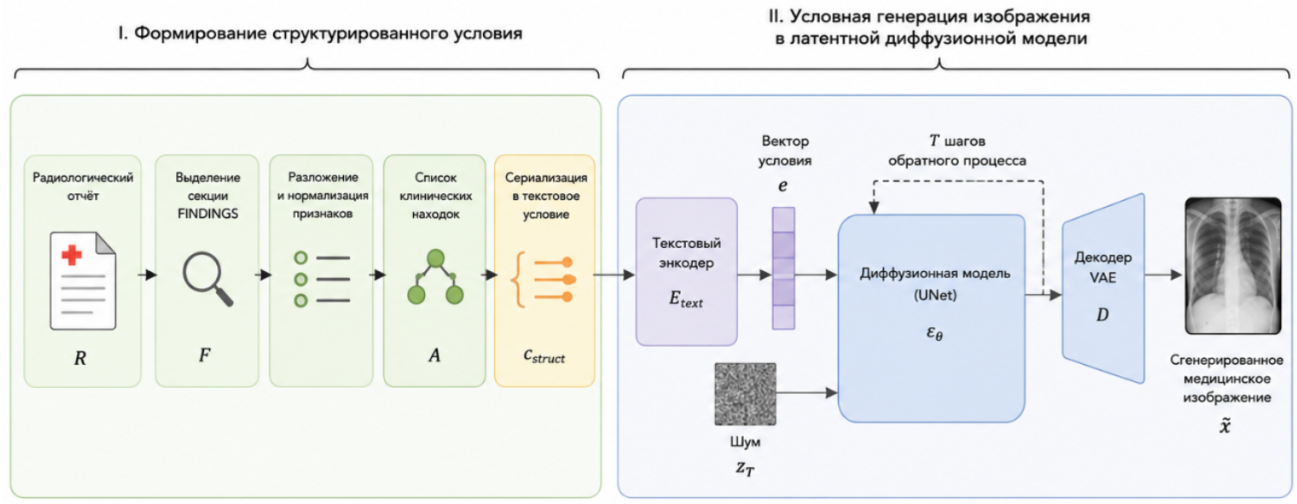


Рисунок 5 — Архитектура программного комплекса условной генерации изображений со структурированным текстовым условием

изображениями и экспериментальная оценка реализованы с использованием специализированных библиотек обработки изображений и машинного обучения. Основные версии программного и аппаратного окружения приведены в Таблице 11.

Таблица 11 — Программное и аппаратное окружение комплекса

Компонент	**Конфигурация**
Python	3.10.13
PyTorch / torchvision	2.10.0 / 0.25.0
Diffusers	0.36.0
Transformers / Tokenizers	5.2.0 / 0.22.2
Accelerate / xFormers	1.12.0 / 0.0.34
Datasets	4.5.0
NumPy / pandas	2.2.6 / 2.3.3
pydicom / Albumentations	3.0.1 / 2.0.8
scikit-learn / scikit-image	1.7.2 / 0.25.2
Собственный пакет	`mimic-sd15-cxr 0.2.0`
Stable Diffusion	версия 1.5
Длина текстового входа	77 токенов
CUDA runtime / cuDNN	12.8 / 9.10.2.21
GPU	4 × NVIDIA A100 80 GB
Операционная система	Ubuntu 22.04

Гиперпараметры обучения и генерации не являются частью архитектуры разработанного комплекса и приводятся вместе с экспериментальным протоколом в Главе 3. Для сравниваемых вариантов используется одна

программная реализация Stable Diffusion; изменяется только представление текстового условия. Параметры генерации и процедуры расчёта метрик также задаются одинаково для обеих ветвей.

После генерации синтетических изображений результаты передаются в общий блок экспериментальной оценки. Он включает расчёт стандартных метрик качества генерации, downstream-оценку практической полезности синтетических данных, поиск ближайших обучающих изображений, проверку возможности определения принадлежности изображения к обучающей выборке и оценку риска повторной идентификации пациента. Постановка этих экспериментов и полученные результаты приведены в Главе 3.

Разработанная программная реализация тем самым охватывает весь используемый в работе конвейер: от подготовки исходного текстового описания и формирования структурированного условия до обучения диффузионной модели, генерации изображений и их последующей экспериментальной оценки. При этом структурированное условие вводится как отдельный этап подготовки данных и не требует изменения внутренних компонентов Stable Diffusion.

2.6 Выводы ко второй главе

Во второй главе разработан метод условной генерации изображений со структурированным текстовым условием и рассмотрена его реализация на примере медицинских данных. В прикладной реализации исходное радиологическое описание преобразуется в структурированное условие в виде списка клинических находок. Основные результаты главы состоят в следующем:

1. Формализовано представление радиологического отчёта и клинической находки. Структурированное условие представлено списком элементов, содержащих тип визуального признака, статус, анатомическую локализацию и степень выраженности.
2. Разработан алгоритм формирования списка клинических находок, включающий выделение секции *FINDINGS*, нормализацию текста, извлечение клинически значимых сущностей и атрибутов, обработку отрицаний, фильтрацию нерелевантных фрагментов, объединение дубликатов и разрешение потенциальных противоречий.
3. Определена сериализация структурированного условия в формате `finding | location | presence | severity`. Отрицательные и

нормальные утверждения учитываются при семантическом разборе, но не включаются как положительные элементы конечного условия; неопределённые положительные находки сохраняются со статусом `probable`.

4. Сформулированы и доказаны семь лемм и три теоремы, декомпозирующие вычислительные и пространственные оценки разработанного метода. Для алгоритма формирования списка получена оценка вычислительной сложности $\mathcal{O}(L + l\,d + C)$, которая при фиксированных параметрах схемы сводится к $\mathcal{O}(L)$. Пространственная сложность хранения списка из m находок оценена как $\mathcal{O}(ma)$ и при ограниченной сверху длине представления одной находки составляет $\mathcal{O}(m)$.
5. Показано, что использование списка находок вместо исходной секции *FINDINGS* не изменяет асимптотическую сложность одного шага обучения и генерации Stable Diffusion при одинаковой максимальной длине текстового входа. Архитектура VAE, U-Net, текстового энкодера и механизм cross-attention при этом не изменяются.
6. Разработана архитектура программного комплекса, объединяющего подготовку радиологических отчётов, формирование двух вариантов текстового условия, обучение и применение Stable Diffusion, а также средства экспериментальной оценки.

Глава 3. Экспериментальная оценка метода условной генерации изображений со структурированным текстовым условием на медицинских данных

3.1 Данные и организация экспериментального исследования

Третья глава посвящена экспериментальной проверке разработанного метода условной генерации изображений со структурированным текстовым условием на медицинских данных. Основной эксперимент строится как прямое сравнение двух вариантов text-to-image генерации на одном и том же наборе рентгенограмм: базовой модели, использующей текстовое условие F, и предлагаемой модели, использующей сформированный во второй главе список клинических находок. Условие F формируется процедурой раздела 2.2. Далее для краткости базовый вариант обозначается как *FINDINGS*. В обоих случаях применяется одна и та же архитектура Stable Diffusion 1.5 [24]; меняется только текстовое представление условия. Для обеих моделей использовался единый протокол обучения и одинаковый критерий выбора итогового состояния модели. В процессе обучения качество контролировалось по значению функции потерь на валидационной выборке, не использовавшейся для обновления параметров модели. Такая постановка позволяет отделить эффект структурирования текста от влияния архитектурных изменений.

Экспериментальная проверка включает шесть взаимосвязанных блоков: сравнение качества генерации по MSE, FID и KID; downstream-оценку синтетических данных на реальной выборке CheXpert; поиск ближайших обучающих изображений; проверку возможности определения принадлежности к обучающей выборке; оценку риска повторной идентификации пациента; а также перенос принципа структурированного условия на маммографические изображения VinDr-Mammo. Совокупность этих экспериментов позволяет оценить качество генерации, практическую полезность синтетических данных и связанные с конфиденциальностью риски.

Общая схема эксперимента показана на Рисунке 6. Она объединяет подготовку двух вариантов условия, обучение генеративных моделей и последующие блоки количественной, downstream- и privacy-оценки.

Рисунок 6 — Схема экспериментального протокола сравнения исходного и структурированного текстового условия.

3.1.1 Используемые наборы данных

Основная экспериментальная база - MIMIC-CXR [42]. Набор содержит рентгенограммы органов грудной клетки в формате DICOM и соответствующие радиологические отчёты. В работе используется исходное разбиение на обучающую, валидационную и тестовую части. Количественные характеристики приведены в Таблице 12.

Таблица 12 — Характеристики экспериментального набора MIMIC-CXR

Характеристика	**Значение**
Тип изображений	Рентгенограммы органов грудной клетки
Формат исходных изображений	DICOM
Число исследований	227 835
Число изображений	377 110
Обучающая выборка	368 960
Валидационная выборка	2 991
Тестовая выборка	5 159

Для каждого изображения основной выборки формируются два варианта текстового условия. В базовой ветви используется очищенная секция *FINDINGS*; в предлагаемой ветви она преобразуется в сериализованный список клинических находок по алгоритму раздела 2.2. При этом визуальная часть, разбиение данных и архитектура генератора совпадают.

Для независимой оценки полезности синтетических рентгенограмм

применяется CheXpert [55]. Генеративные модели на CheXpert не дообучаются: реальная часть этого набора используется только на этапе оценки классификаторов, обученных на синтетических изображениях. Для дополнительного маммографического эксперимента используется VinDr-Mammo [46]. Назначение наборов суммировано в Таблице 13.

Таблица 13 — Назначение наборов данных в экспериментальном исследовании

Набор	**Тип данных**	**Роль в эксперименте**
MIMIC-CXR	CXR и радиологические отчёты	Обучение двух генеративных моделей; MSE/FID/KID; nearest-train; membership inference; re-identification
CheXpert	Размеченные CXR	Независимая downstream-оценка классификаторов
VinDr-Mammo	Маммограммы и структурированные атрибуты	Проверка переносимости принципа структурированного условия и синтетической аугментации

3.1.2 Подготовка изображений и текстовых условий

Изображения MIMIC-CXR извлекаются из DICOM-файлов и приводятся к формату, совместимому с VAE Stable Diffusion. Один и тот же конвейер визуальной предобработки используется в обеих ветвях. Текстовый конвейер начинается с полного радиологического отчёта и состоит из двух этапов, подробно описанных во второй главе: выделения релевантного содержимого FINDINGS и формирования компактного списка находок. Подготовленные данные сохраняются в CSV, что позволяет отделить однократную подготовку текста от многократных эпох обучения генеративной модели.

В использованной Stable Diffusion 1.5 длина входа текстового энкодера ограничена 77 токенами. Поэтому для базовой и предлагаемой ветвей применяется одинаковая политика токенизации и обрезки. Структурированное условие, как правило, компактнее свободного FINDINGS: из него удаляются отрицательные, нормальные, исторические, технические и дублирующиеся фрагменты. Благодаря этому снижается риск того, что клинически значимая часть условия окажется за границей доступного текстового контекста. Распределение длины двух вариантов условия является дополнительной характеристикой эффекта структурирования и должно быть рассчитано тем же токенизатором, который используется при обучении Stable Diffusion.

Для количественной оценки компактности текстового условия была измерена длина входных последовательностей с использованием того же CLIP-токенизатора Stable Diffusion 1.5, который применялся при обучении модели. Длина последовательности определялась до её обрезки до максимального контекста текстового энкодера, равного 77 токенам. Полученные статистические характеристики длины исходной секции FINDINGS и структурированного списка находок приведены в Таблице 14.

Таблица 14 — Характеристики длины текстового условия до обрезки

Условие	**Среднее**	**Медиана**	P_{95}	**Максимум**	**Доля** > 77**, %**
FINDINGS	74,48	69	134	549	38,16
Список находок	34,88	29	87	243	8,03

Полученные результаты показывают, что структурированное представление существенно компактнее исходного текстового условия. Средняя длина условия уменьшается более чем в два раза, а медианная длина снижается с уровня, близкого к предельному размеру контекста текстового энкодера, до значения, существенно меньшего данного ограничения. Аналогичная тенденция наблюдается и для верхней части распределения: значение 95-го процентиля для списка находок значительно ниже, чем для исходной секции FINDINGS.

Наиболее существенное различие связано с долей последовательностей, превышающих ограничение в 77 токенов. При использовании исходной секции FINDINGS данное ограничение превышает более трети условий, тогда как для структурированного списка находок таких случаев менее одной десятой выборки. Таким образом, структурирование текста существенно уменьшает число условий, для которых при передаче в текстовый энкодер требуется обрезка входной последовательности, и тем самым снижает вероятность потери части информации из-за ограничения длины контекста.

Распределения длины двух вариантов текстового условия представлены на Рисунке 7.

3.1.3 Параметры обучения, генерации и воспроизводимость

Оба экземпляра генератора основаны на Stable Diffusion 1.5 и обучаются с одинаковой конфигурацией. Фиксация одинаковых гиперпараметров

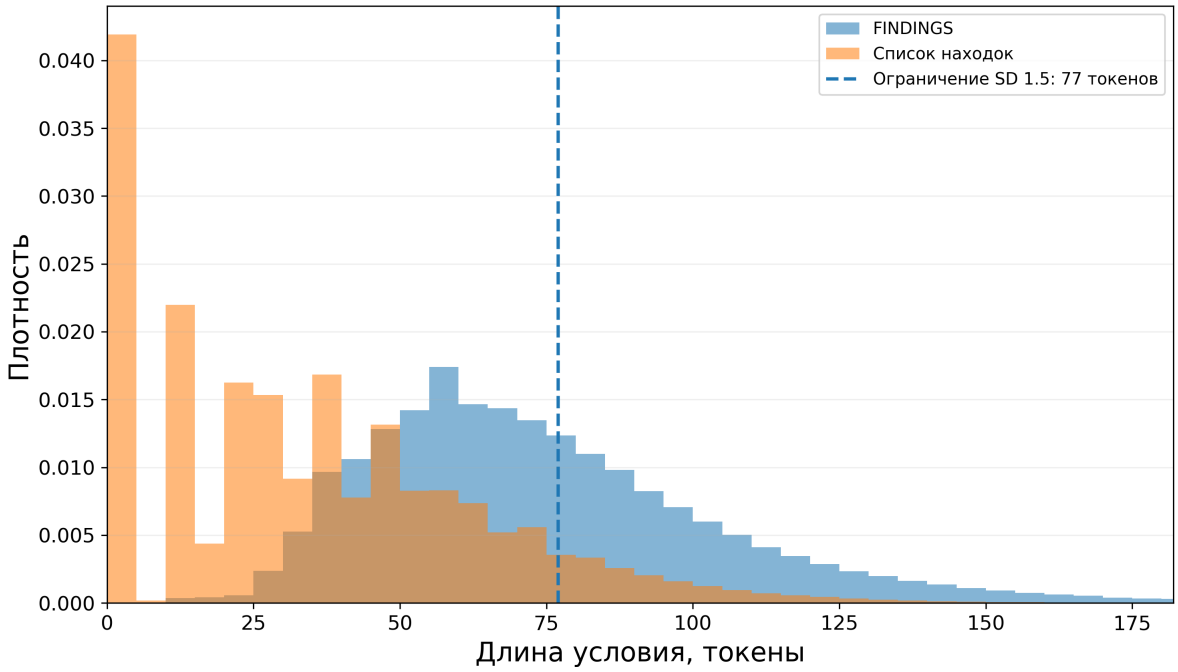


Рисунок 7 — Распределение длины текстового условия для исходной секции FINDINGS и структурированного списка находок. Вертикальной пунктирной линией обозначено ограничение текстового энкодера Stable Diffusion 1.5, равное 77 токенам.

принципиальна: различие между результатами должно определяться формой условия, а не разрешением изображения, оптимизатором или параметрами сэмплирования.

Параметры обучения и генерации сведены в таблицу 15.

Для FID и KID используются стандартные признаки Inception-v3 [69, 58, 59]; SSIM рассчитывается в соответствии с [60]. ROC AUC и PR AUC интерпретируются согласно [61, 62].

3.2 Оценка качества генерации

В первом эксперименте две модели оцениваются на одной тестовой выборке MIMIC-CXR. Для каждого тестового условия формируется синтетическое изображение, после чего вычисляются MSE, FID и KID.

Результаты приведены в Таблице 16.

При структурированном условии FID снижается с 64.156 до 61.245, то есть на 4.5%, а KID - с 0.0185 до 0.0159, то есть на 14.1%. MSE, напротив, незначительно увеличивается с 0.0901 до 0.0921. Это не является противоречием: MSE характеризует попиксельное сходство конкретной пары,

Таблица 15 — Параметры обучения и генерации двух сравниваемых моделей

Параметр	**Фактическое значение**
Архитектура	Stable Diffusion 1.5
Максимальная длина текстового входа	77 токенов
Python	3.10.13
PyTorch / Diffusers / Transformers	2.10.0 / 0.36.0 / 5.2.0
Формат подготовленных условий	CSV
Разрешение обучающих изображений	256×256 пикселей
Batch size	512
Gradient accumulation	1
Оптимизатор	AdamW
Learning rate и scheduler обучения	$5 \cdot 10^{-5}$; линейный warmup в течение первых 100 шагов, далее постоянный LR
Обучаемые компоненты (U-Net/text encoder/VAE)	обучается U-Net; text encoder и VAE заморожены
Mixed precision	bf16
Random seed обучения	42
Scheduler генерации	DDIM
Число шагов денойзинга	30
Classifier-free guidance scale	4.0
Random seed генерации	начальный seed 42
GPU и объём VRAM	NVIDIA A100 80 GB

Таблица 16 — Сравнение моделей по метрикам качества генерации

Условие	**MSE**	**FID**	**KID**
FINDINGS	**0.0901**	64.156	0.0185
Список находок	0.0921	**61.245**	**0.0159**

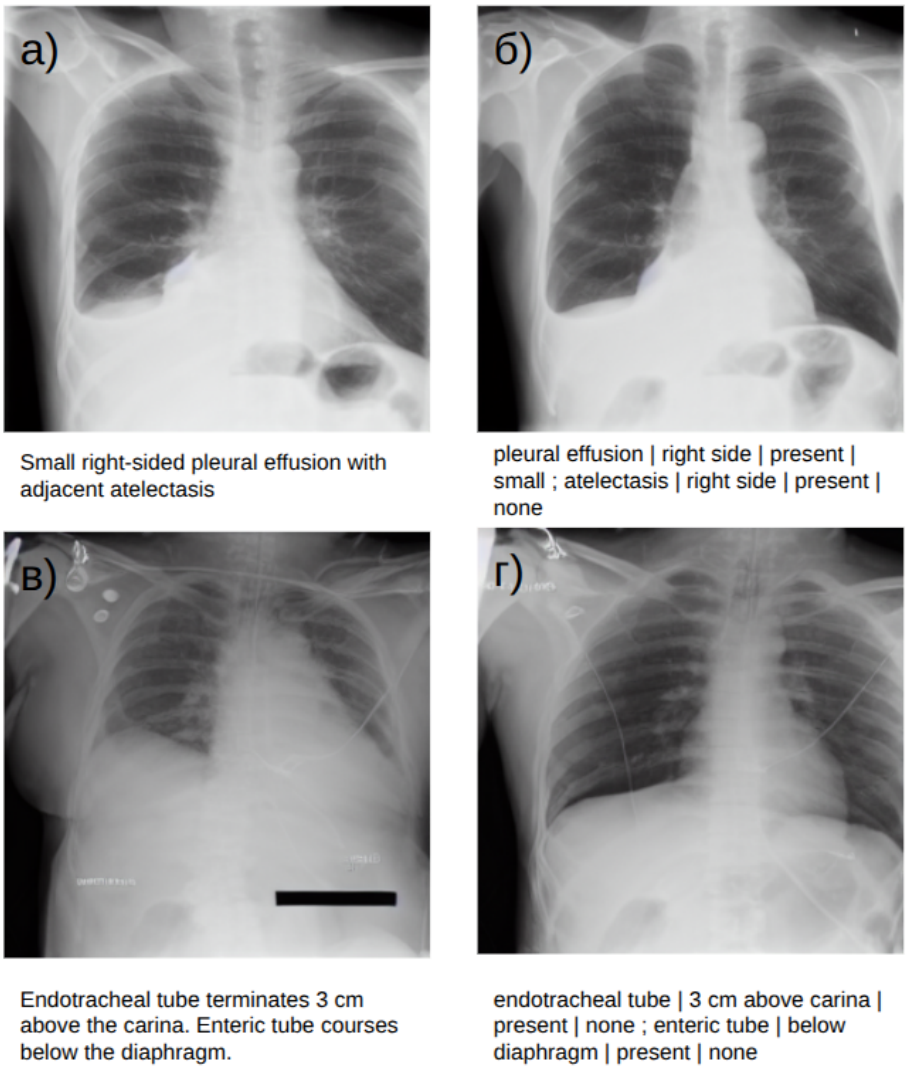


Рисунок 8 — Примеры синтетических рентгенограмм при двух способах задания текстового условия.

тогда как FID и KID сопоставляют распределения признаков реальных и синтетических выборок. Для стохастической генерации задача не состоит в реконструкции конкретной рентгенограммы пиксель в пиксель.

Примеры изображений, полученных двумя моделями для сопоставимых условий, приведены на Рисунке 8.

Таким образом, структурирование условия улучшает оба распределительных показателя и не приводит к принципиальному ухудшению попиксельной метрики. Далее распределительное сравнение дополняется проверкой практической полезности синтетических данных.

3.3 Downstream-оценка практической полезности синтетических данных

Downstream-эксперимент построен по схеме *train on synthetic, test on real*. Для каждого генератора создаётся отдельный синтетический обучающий набор, затем на нём обучается классификатор, который оценивается на одной и той же независимой реальной выборке CheXpert [55]. Такой протокол проверяет, сохраняются ли в синтетических изображениях признаки, которые переносятся на реальные CXR.

В обеих ветвях downstream-эксперимента использовался классификатор EfficientNet-B3 с весами, предварительно обученными на ImageNet. Каждый синтетический набор содержал 2936 изображений и при random seed 42 разделялся на обучающую, валидационную и внутреннюю тестовую части размером 2348, 294 и 294 изображения соответственно.

Решалась многометочная задача классификации по шести классам: *No Finding*, *Cardiomegaly*, *Edema*, *Pleural Effusion*, *Atelectasis* и *Pneumothorax*. Для обучения использовался AdamW с learning rate 10^{-4}, weight decay 10^{-4} и batch size 256. Максимальное число эпох составляло 30. Функцией потерь являлась BCEWithLogitsLoss без дополнительного взвешивания положительных классов. Итоговый checkpoint выбирался по минимальному значению ошибки на валидационной выборке. Обе модели оценивались на одной и той же независимой реальной выборке CheXpert. Результаты приведены в Таблице 17.

Таблица 17 — Downstream-оценка полезности синтетических данных на CheXpert

Источник синтетики	**Macro ROC AUC**	**Macro PR AUC**	**Macro F1**
FINDINGS	0.6205	0.2145	0.1507
Список находок	**0.6252**	**0.2172**	**0.1637**

Классификатор, обученный на изображениях со структурированным условием, показывает небольшое, но однонаправленное преимущество по всем трём метрикам. Наиболее заметное относительное изменение наблюдается для Macro F1: рост с 0.1507 до 0.1637 соответствует 8.6%. Macro ROC AUC увеличивается на 0.0047, Macro PR AUC - на 0.0027. Без повторных запусков или доверительных интервалов эти небольшие различия следует трактовать

как наблюдаемую тенденцию, а не как доказательство статистически значимого превосходства.

В совокупности с FID и KID этот эксперимент показывает, что улучшение распределительных метрик не сопровождается потерей downstream-полезности.

3.4 Оценка сходства синтетических и обучающих изображений

Следующий эксперимент проверяет, не являются ли синтетические изображения практически точными копиями отдельных объектов обучающей выборки. Для каждого синтетического изображения находится ближайший train-объект; в качестве контроля та же процедура выполняется для независимых реальных test-изображений. Сравниваются направления synthetic→train и test→train. Подобные проверки используются при анализе запоминания генеративными моделями [64, 65].

Близость оценивалась по косинусному расстоянию в признаковом пространстве и SSIM [60]. В качестве feature encoder использовалась DenseNet121 с весами, предварительно обученными на ImageNet; классификационный слой заменялся тождественным отображением. Размерность получаемого признакового представления составляла 1024. Перед извлечением признаков изображения преобразовывались в RGB, масштабировались до 224×224 пикселей и нормализовались с использованием статистик ImageNet.

Для каждой ветви train-галерея содержала 20000 реальных изображений, контрольная test-подвыборка - 1000 реальных изображений, а синтетическая выборка - 2936 изображений. Поиск ближайшего соседа выполнялся по косинусному расстоянию с помощью `NearestNeighbors` библиотеки scikit-learn при параметрах `n_neighbors=1`, `metric=cosine` и `algorithm=brute`. Для вычисления SSIM соответствующие пары изображений дополнительно приводились к одноканальному представлению. Результаты приведены в Таблице 18.

Для обоих вариантов генерации синтетические изображения в среднем находятся дальше от ближайшего train-объекта по косинусному расстоянию и имеют меньший SSIM, чем независимые реальные тестовые изображения. Следовательно, в рамках использованной подвыборки и выбранного признакового представления признаки прямого воспроизведения отдельных

Таблица 18 — Результаты поиска ближайших обучающих изображений

Условие	**Косинусное расстояние, synthetic / test**	**SSIM, synthetic / test**
FINDINGS	0.0987 / 0.0822	0.445 / 0.509
Список находок	0.1017 / 0.0807	0.436 / 0.508

обучающих изображений не выявлены.

Помимо средних значений информативны полные распределения расстояний, которые приведены на Рисунке 9 и строятся по индивидуальным результатам поиска ближайших соседей.

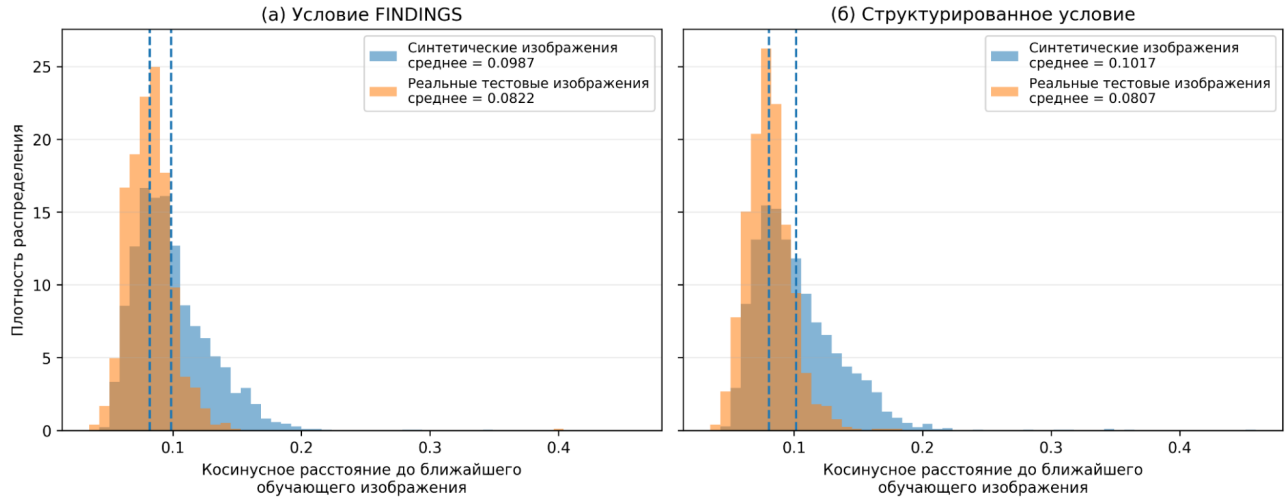


Рисунок 9 — Распределения косинусного расстояния до ближайшего обучающего изображения для синтетических и реальных тестовых изображений.

3.5 Оценка возможности определения принадлежности к обучающей выборке

В отдельном эксперименте проверяется, можно ли отличить реальные train- и real test-изображения по расстоянию до синтетического набора. Для каждого реального изображения находится ближайший synthetic-объект, а полученная мера близости используется как score бинарного классификатора “train/test”. Подобная постановка относится к классу membership inference attacks [63, 64].

Использовалось то же признаковое представление DenseNet121 размерности 1024, что и в эксперименте поиска ближайших обучающих

изображений. Для каждого варианта генерации формировались две подвыборки по 1000 реальных изображений: изображения обучающей выборки и независимые тестовые изображения. В качестве референтной синтетической галереи использовались 2936 изображений.

Для каждого реального изображения находился ближайший синтетический объект. При построении ROC изображения обучающей выборки рассматривались как положительный класс, а тестовые изображения - как отрицательный. Для косинусного расстояния и MSE меньшие значения соответствуют большей близости, поэтому в качестве score использовались $-d_{\cos}$ и $-\mathrm{MSE}$. Для SSIM большее значение соответствует большей близости, поэтому использовалось непосредственно значение SSIM. ROC AUC вычислялся с помощью функции `roc_auc_score` библиотеки scikit-learn.

Результаты приведены в Таблице 19.

Таблица 19 — Определение принадлежности к обучающей выборке по близости к синтетическим данным

	Косинусное расстояние		**SSIM**		**MSE**	
Условие	**train/test**	**AUC**	**train/test**	**AUC**	**train/test**	**AUC**
FINDINGS	0.1011 / 0.1024	0.5149	0.4900 / 0.4897	0.5002	0.0621 / 0.0625	0.5082
Список находок	0.1043 / 0.1020	0.4641	0.4874 / 0.4888	0.4924	0.0625 / 0.0638	0.5068

Все значения ROC AUC находятся близко к 0.5. Максимальное абсолютное отклонение от случайного уровня составляет 0.0359 (AUC=0.4641). Средние значения мер близости для train и test также практически совпадают. Поэтому выбранные scores не позволяют надёжно восстановить факт участия изображения в обучении генератора.

3.6 Оценка риска повторной идентификации пациента

Третий privacy-эксперимент проверяет более специфичный риск: сохранение в синтетическом изображении признаков, позволяющих сопоставить его с конкретным пациентом. Используется специализированный энкодер на основе ResNet50 [70], обученный методом supervised contrastive learning [71]. Синтетическое изображение служит запросом, а реальные рентгенограммы образуют галерею. Качество поиска оценивается по Top-1, Top-5 и mAP.

Перед синтетическими запросами выполнялся положительный контроль real→real. Он использовался для проверки того, что выбранное признаковое пространство действительно содержит устойчивый patient-specific сигнал.

Энкодер повторной идентификации был построен на основе ResNet50 с весами, предварительно обученными на ImageNet. Размерность итогового embedding составляла 512, а размерность projection head при contrastive training - 128. Использовался dropout 0.1. Входные изображения масштабировались до 256×256 пикселей, полученные embedding нормировались по L_2-норме, а поиск выполнялся по косинусному сходству.

Для обучения энкодера использовалась выборка из 100000 рентгенограмм 15890 пациентов. Разбиение выполнялось по пациентам: обучающая часть содержала 69682 изображения 11123 пациентов, валидационная - 15306 изображений 2384 пациентов, тестовая - 15012 изображений 2383 пациентов. Таким образом, изображения одного пациента не могли одновременно присутствовать в разных частях выборки.

Обучение выполнялось с использованием supervised contrastive loss при temperature 0.07. Один batch формировался из 128 пациентов по два изображения каждого пациента, то есть эффективный batch size составлял 256. Использовался AdamW с learning rate 10^{-4} и weight decay 10^{-5}; обучение выполнялось в течение 100 эпох. Итоговый checkpoint выбирался по максимальному значению mAP на валидационной выборке.

Результаты показаны в Таблице 20.

Таблица 20 — Результаты оценки риска повторной идентификации пациента

Тип запроса	**Top-1**	**Top-5**	**mAP**
Real → real gallery	0.7700	0.9110	0.7040
Synthetic, FINDINGS → real gallery	0.0135	0.0296	0.0155
Synthetic, список находок → real gallery	**0.0091**	**0.0236**	**0.0128**

Положительный контроль показывает высокую идентифицируемость реальных рентгенограмм: Top-1=0.7700, Top-5=0.9110, mAP=0.7040. При использовании синтетических запросов показатели снижаются более чем на 96% относительно поиска real→real. Для структурированного условия дополнительно наблюдается снижение Top-1 с 0.0135 до 0.0091, Top-5 с 0.0296 до 0.0236 и mAP с 0.0155 до 0.0128 относительно FINDINGS.

Для оценки устойчивости направления изменений применялся парный бутстрэп [72]. Сопоставление выполнялось для 15012 пар синтетических запросов, соответствующих одним и тем же исходным объектам. Единицей ресэмплинга являлся отдельный запрос. Использовалось $B = 1000$ бутстрэп-повторов при random seed 42; 95%-е доверительные интервалы рассчитывались процентильным методом. Для каждой метрики анализировалась разность “список находок минус FINDINGS”. Во всех трёх случаях доверительный интервал не пересекает нулевое значение, что подтверждает устойчивость наблюдаемого направления снижения показателей повторной идентификации при использовании структурированного условия в рамках выбранной процедуры оценки.

3.7 Перенос метода на маммографические изображения

Дополнительный эксперимент проводится на VinDr-Mammo [46]. Здесь свободный отчёт не используется: условие сразу строится из структурированных атрибутов исследования - проекции, стороны, категории плотности молочной железы, BI-RADS и типов находок. Поэтому эксперимент проверяет переносимость общего принципа структурированного задания условия, а не повторяет алгоритм извлечения FINDINGS.

Такая постановка содержательно связана с опубликованной работой по privacy-preserving synthetic mammograms [4]; VinDr-Mammo также использовался в работе по многопроекционной классификации MamT^4 [73].

Сравниваются пять составов обучающей выборки классификатора: только реальные данные; реальные данные с добавлением синтетической части размером 25%, 50% и 100% от размера обучающей выборки; а также обучение только на синтетических данных. Во всех случаях оценка проводится на одной и той же независимой реальной test-выборке.

Обучающая часть VinDr-Mammo содержала 15954 изображения, соответствующих 4000 исследованиям, а независимая тестовая часть - 3993 изображения 1000 исследований. Для обучения генеративной модели выборка была сокращена для балансировки.

В итоге объёмы синтетических дополнений рассчитывались относительно размера обучающей выборки генеративной модели. Для вариантов 25%, 50% и 100% использовались соответственно 1024, 2048 и 4095 синтетических

изображений. Результаты приведены в Таблице 21.

Таблица 21 — Классификация на VinDr-Mammo при различных составах обучающей выборки

Обучающая выборка	**Synthetic**	**ROC AUC**	**Accuracy**	**Precision**	**Recall**	**Specificity**
Реальные	0%	0.9406	0.9918	0.7000	**0.6087**	0.9967
Реальные + синтетические	25%	0.9330	0.9896	0.6000	0.5217	0.9956
Реальные + синтетические	50%	0.9252	0.9904	0.6222	**0.6087**	0.9953
Реальные + синтетические	100%	**0.9494**	**0.9921**	**0.7742**	0.5217	**0.9981**
Только синтетические	100%	0.8138	0.9738	0.1429	0.2174	0.9834

Зависимость качества от доли синтетических данных немонотонна. Добавление 25% и 50% синтетики не улучшает базовый ROC AUC. При равном объёме real и synthetic ROC AUC увеличивается с 0.9406 до 0.9494, Precision - с 0.7000 до 0.7742, также немного возрастают Accuracy и Specificity. Одновременно Recall снижается с 0.6087 до 0.5217, поэтому результат нельзя трактовать как одновременное улучшение всех характеристик классификатора.

Обучение только на синтетических данных приводит к заметному снижению ROC AUC, Precision и Recall. Следовательно, синтетические маммограммы в данной постановке полезны прежде всего как дополнение к реальной выборке, но не как её полная замена.

3.8 Обобщение результатов и ограничения исследования

Сводные результаты экспериментов приведены в Таблице 22. В неё включены только ключевые численные результаты, используемые далее при формулировании выводов главы.

Полученные результаты показывают, что структурированное условие обеспечивает снижение FID на 4.5% и KID на 14.1%, а Macro F1 в downstream-задаче увеличивается на 8.6%. Анализ ближайших изображений не выявляет практически точного воспроизведения train-объектов; membership inference остаётся около случайного уровня; re-identification существенно слабее поиска между реальными изображениями. На VinDr-Mammo наилучший ROC AUC среди смешанных наборов получен при равном объёме real и synthetic.

При интерпретации результатов необходимо учитывать ограничения. Во-первых, MSE, FID и KID не являются прямой проверкой клинической корректности каждой генерации. Во-вторых, downstream-оценка зависит от

Таблица 22 — Основные результаты экспериментальной главы

Эксперимент	**Ключевой результат**	**Интерпретация**
Качество генерации	FID 64.156 → 61.245; KID 0.0185 → 0.0159	Улучшение распределительных метрик
Downstream CheXpert	Macro F1 0.1507 → 0.1637	Прирост 8.6%, полезность синтетики сохраняется
Nearest train	Synthetic дальше от train, чем independent test	Прямые копии не выявлены выбранным методом
Membership inference	ROC AUC 0.4641–0.5149 (cosine) и около 0.5 для остальных мер	Выраженный membership-сигнал не выявлен
Re-identification	Top-1: 0.7700 real→real; 0.0135/0.0091 synthetic	Более чем 96% снижение относительно real-контроля
VinDr-Mammo	ROC AUC 0.9406 → 0.9494 при real+100% synthetic	Синтетика полезна как дополнение, но не замена real

выбранного классификатора, набора классов и протокола обучения. В-третьих, privacy-эксперименты зависят от feature encoder, меры близости, состава галереи и возможностей атакующего; отсутствие обнаруженного сигнала не является доказательством абсолютной конфиденциальности. В-четвёртых, эксперимент VinDr-Mammo характеризуется выраженным дисбалансом классов, поэтому высокая Accuracy должна анализироваться совместно с Precision, Recall и Specificity. Наконец, основной прямой эксперимент FINDINGS против списка находок проведён на одной медицинской модальности и одной базовой архитектуре Stable Diffusion 1.5.

Таким образом, проведённая экспериментальная проверка подтверждает практическую применимость разработанного метода структурированного задания условия. При неизменной архитектуре генератора список клинических находок улучшает распределительные метрики и downstream-показатели; в рассмотренных privacy-постановках увеличение рисков не обнаружено; принцип структурированного условия переносится на маммографические изображения.

3.9 Выводы к третьей главе

В третьей главе выполнена экспериментальная проверка разработанного метода на рентгенограммах органов грудной клетки и дополнительная проверка переносимости принципа структурированного условия на маммографические изображения. Получены следующие результаты.

1. При одинаковой архитектуре Stable Diffusion 1.5 переход от секции *FINDINGS* к списку клинических находок снизил FID с 64.156 до 61.245, то есть приблизительно на 4.5%, и KID с 0.0185 до 0.0159, то есть приблизительно на 14.1%. MSE при этом изменился с 0.0901 до 0.0921, что подтверждает необходимость совместного анализа нескольких метрик.
2. В downstream-эксперименте на независимых реальных данных CheXpert классификатор, обученный на синтетических изображениях со структурированным условием, получил Macro F1=0.1637 против 0.1507 для FINDINGS, что соответствует относительному увеличению приблизительно на 8.6%. Macro ROC AUC и Macro PR AUC также немного увеличились.
3. Поиск ближайших обучающих изображений не выявил систематической близости синтетических объектов к train-выборке, превышающей контроль real test→train. В рамках использованного признакового представления признаки практически точного воспроизведения обучающих изображений не обнаружены.
4. В membership inference значения ROC AUC для косинусного расстояния, SSIM и MSE находились в диапазоне, близком к случайному уровню 0.5. Следовательно, выбранные меры близости не позволили надёжно определить, входило ли реальное изображение в обучающую выборку генератора.
5. В эксперименте повторной идентификации положительный контроль real→real подтвердил наличие patient-specific сигнала (Top-1=0.7700, Top-5=0.9110, mAP=0.7040). Для синтетических запросов показатели снизились более чем на 96%; для списка находок получены Top-1=0.0091, Top-5=0.0236 и mAP=0.0128.
6. На VinDr-Mammo зависимость качества от объёма синтетической части оказалась немонотонной. При добавлении синтетических данных ROC AUC увеличился с 0.9406 до 0.9494, а Precision - с 0.7000 до 0.7742,

при одновременном снижении Recall с 0.6087 до 0.5217. Обучение только на синтетических данных существенно уступило обучению на реальных данных, поэтому синтетика рассматривается как дополнение, а не как замена реальной выборки.

Совокупность экспериментов показывает, что структурированное задание условия улучшает распределительные показатели генерации и сохраняет практическую полезность синтетических данных. Проведённые privacy-эксперименты не выявили увеличения рассматриваемых рисков, однако их результаты относятся к конкретным признаковым представлениям и сценариям атак и не являются универсальной гарантией конфиденциальности.

ЗАКЛЮЧЕНИЕ

В диссертационной работе разработан и экспериментально проверен метод условной генерации изображений на основе диффузионных моделей со структурированным заданием условия, формируемым из текстового описания. Разработка и экспериментальная валидация метода выполнены на примере медицинских данных. Предложенный подход обеспечивает преобразование неструктурированного текстового описания в компактное структурированное представление, пригодное для использования в латентной диффузионной модели.

Основные результаты диссертационной работы заключаются в следующем:

1. Разработан метод формирования структурированного условия для диффузионной генерации изображений на основе разложения неструктурированного текстового описания на семантически значимые компоненты, исследованный на примере медицинских данных.
2. Разработан алгоритм формирования списка клинических находок, включающий выделение релевантной части радиологического отчета, нормализацию клинических формулировок, обработку отрицаний и неопределенности и сериализацию результата в единый формат структурированного текстового условия.
3. Сформулированы и доказаны семь лемм и три теоремы, на основе которых получены аналитические оценки вычислительной и пространственной сложности разработанных алгоритмов. Показано, что вычислительная сложность формирования списка находок линейна относительно длины обрабатываемого текста при фиксированных параметрах схемы, а пространственная сложность хранения структурированного условия линейна относительно числа выделенных находок при ограниченной длине одной записи. Замена секции FINDINGS структурированным списком находок не изменяет асимптотическую сложность обучения и генерации Stable Diffusion.
4. Разработаны архитектура и программная реализация комплекса условной генерации изображений со структурированным текстовым

условием, включающего подготовку текстовых описаний, формирование структурированного условия, обучение и применение Stable Diffusion 1.5, а также средства экспериментальной оценки синтетических данных на примере медицинских данных.

5. Проведена экспериментальная валидация метода на рентгенограммах органов грудной клетки. Использование структурированного условия снизило FID с 64.156 до 61.245 (на 4.5%) и KID с 0.0185 до 0.0159 (на 14.1%), а значение Macro F1 в downstream-задаче увеличилось с 0.1507 до 0.1637 (на 8.6%). Дополнительные эксперименты не выявили признаков практически точного воспроизведения обучающих изображений; значения ROC AUC определения принадлежности к обучающей выборке оставались близкими к 0.5, а показатели повторной идентификации по синтетическим изображениям были существенно ниже, чем при поиске между реальными изображениями.
6. Проверена переносимость подхода на маммографические изображения набора VinDr-Mammo. При добавлении синтетических данных ROC AUC классификатора увеличился с 0.9406 до 0.9494, а Precision - с 0.7000 до 0.7742. При этом обучение только на синтетических данных приводило к заметному снижению качества, что показывает целесообразность их использования прежде всего как дополнения к реальной обучающей выборке.

Таким образом, поставленная цель достигнута: разработан и программно реализован метод условной генерации изображений со структурированным текстовым условием, получены оценки сложности входящих в него алгоритмов и выполнена экспериментальная валидация метода на медицинских данных. Полученные результаты показывают, что структурирование текстового условия позволяет улучшить ряд метрик качества генерации и сохранить практическую полезность синтетических изображений.

Список литературы

[1] *Ushakov E. et al. Endonet: A model for the automatic calculation of H-score on histological slides //Informatics. – MDPI, 2023. – Т. 10. – №. 4. – С. 90.*

[2] *Ushakov E. et al. Optimizing the Mammography AI Pipeline: From Data Filtering to Vision-Language Models //Informatics. – MDPI, 2026. – Т. 13. – №. 8. – С. 125.*

[3] *Litvinov A. et al. Clinically aware learning: Ordinal loss improves medical image classifiers //Journal of Clinical Medicine. – 2026. – Т. 15. – №. 1. – С. 365.*

[4] *Shodiev D. et al. Privacy-preserving synthetic mammograms: a generative model approach to privacy-preserving breast imaging datasets //Informatics. – MDPI, 2025. – Т. 12. – №. 4. – С. 112.*

[5] *Ackley D. H., Hinton G. E., Sejnowski T. J. A learning algorithm for Boltzmann machines //Cognitive science. – 1985. – Т. 9. – №. 1. – С. 147-169.*

[6] *Hinton G. E. et al. The"wake-sleep"algorithm for unsupervised neural networks //Science. – 1995. – Т. 268. – №. 5214. – С. 1158-1161.*

[7] *Hinton G. E., Osindero S., Teh Y. W. A fast learning algorithm for deep belief nets //Neural computation. – 2006. – Т. 18. – №. 7. – С. 1527-1554.*

[8] *Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks //science. – 2006. – Т. 313. – №. 5786. – С. 504-507.*

[9] *Kingma D. P., Welling M. Auto-encoding variational bayes //arXiv preprint arXiv:1312.6114. – 2013.*

[10] *Goodfellow I. J. et al. Generative adversarial nets //Advances in neural information processing systems. – 2014. – Т. 27.*

[11] *Van Den Oord A., Kalchbrenner N., Kavukcuoglu K. Pixel recurrent neural networks //International conference on machine learning. – PMLR, 2016. – С. 1747-1756.*

[12] *Van den Oord A. et al. Conditional image generation with pixelcnn decoders //Advances in neural information processing systems. – 2016. – T. 29.*

[13] *Dinh L., Sohl-Dickstein J., Bengio S. Density estimation using real nvp //arXiv preprint arXiv:1605.08803. – 2016.*

[14] *Kingma D. P., Dhariwal P. Glow: Generative flow with invertible 1x1 convolutions //Advances in neural information processing systems. – 2018. – T. 31.*

[15] *Radford A., Metz L., Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks //arXiv preprint arXiv:1511.06434. – 2015.*

[16] *Arjovsky M., Chintala S., Bottou L. Wasserstein generative adversarial networks //International conference on machine learning. – Pmlr, 2017. – C. 214-223.*

[17] *Karras T., Laine S., Aila T. A style-based generator architecture for generative adversarial networks //2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). – IEEE, 2019. – C. 4396-4405.*

[18] *Sohl-Dickstein J. et al. Deep unsupervised learning using nonequilibrium thermodynamics //International conference on machine learning. – pmlr, 2015. – C. 2256-2265.*

[19] *Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models //Advances in neural information processing systems. – 2020. – T. 33. – C. 6840-6851.*

[20] *Dhariwal P., Nichol A. Diffusion models beat gans on image synthesis //Advances in neural information processing systems. – 2021. – T. 34. – C. 8780-8794.*

[21] *Song J., Meng C., Ermon S. Denoising diffusion implicit models //arXiv preprint arXiv:2010.02502. – 2020.*

[22] *Nichol A. et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models //arXiv preprint arXiv:2112.10741. – 2021.*

[23] *Saharia C. et al. Photorealistic text-to-image diffusion models with deep language understanding //Advances in neural information processing systems. – 2022. – T. 35. – C. 36479-36494.*

[24] *Rombach R. et al. High-resolution image synthesis with latent diffusion models //2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). – ieee, 2022. – C. 10674-10685.*

[25] *Song Y. et al. Score-based generative modeling through stochastic differential equations //arXiv preprint arXiv:2011.13456. – 2020.*

[26] *Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation //International Conference on Medical image computing and computer-assisted intervention. – Cham : Springer international publishing, 2015. – C. 234-241.*

[27] *Lu C. et al. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps //Advances in neural information processing systems. – 2022. – T. 35. – C. 5775-5787.*

[28] *Karras T. et al. Elucidating the design space of diffusion-based generative models //Advances in neural information processing systems. – 2022. – T. 35. – C. 26565-26577.*

[29] *Peebles W., Xie S. Scalable diffusion models with transformers //2023 IEEE/CVF International Conference on Computer Vision (ICCV). – IEEE, 2023. – C. 4172-4182.*

[30] *Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – T. 30.*

[31] *Perez E. et al. Film: Visual reasoning with a general conditioning layer //Proceedings of the AAAI conference on artificial intelligence. – 2018. – T. 32. – №. 1.*

[32] *Ho J., Salimans T. Classifier-free diffusion guidance //arXiv preprint arXiv:2207.12598. – 2022.*

[33] *Zhang L., Rao A., Agrawala M. Adding conditional control to text-to-image diffusion models //2023 IEEE/CVF International Conference on Computer Vision (ICCV). – IEEE, 2023. – С. 3813-3824.*

[34] *Mou C. et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models //Proceedings of the AAAI conference on artificial intelligence. – 2024. – Т. 38. – №. 5. – С. 4296-4304.*

[35] *Radford A. et al. Learning transferable visual models from natural language supervision //International conference on machine learning. – PmLR, 2021. – С. 8748-8763.*

[36] *Zhang Y. et al. Contrastive learning of medical visual representations from paired images and text //Machine learning for healthcare conference. – PMLR, 2022. – С. 2-25.*

[37] *Huang S. C. et al. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition //2021 IEEE/CVF International Conference on Computer Vision (ICCV). – IEEE, 2021. – С. 3922-3931.*

[38] *Boecking B. et al. Making the most of text semantics to improve biomedical vision–language processing //European conference on computer vision. – Cham : Springer Nature Switzerland, 2022. – С. 1-21.*

[39] *Wang Z. et al. Medclip: Contrastive learning from unpaired medical images and text //Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. – 2022. – С. 3876-3887.*

[40] *Chambon P. et al. Roentgen: vision-language foundation model for chest x-ray generation //arXiv preprint arXiv:2211.12737. – 2022.*

[41] *Weber T. et al. Cascaded latent diffusion models for high-resolution chest x-ray synthesis //Pacific-Asia conference on knowledge discovery and data mining. – Cham : Springer Nature Switzerland, 2023. – С. 180-191.*

[42] *Johnson A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports //Scientific data. – 2019. – Т. 6. – №. 1. – С. 317.*

[43] *Korkinof D. et al. High-resolution mammogram synthesis using progressive generative adversarial networks //arXiv preprint arXiv:1807.03401. – 2018.*

[44] *Wu E. et al. Conditional infilling GANs for data augmentation in mammogram classification //International Workshop on Reconstruction and Analysis of Moving Body Organs. – Cham : Springer International Publishing, 2018. – C. 98-106.*

[45] *Skipina M. et al. MAMBO: High-resolution generative approach for mammography images //arXiv preprint arXiv:2506.08677. – 2025.*

[46] *Nguyen H. T. et al. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography //Scientific Data. – 2023. – T. 10. – №. 1. – C. 277.*

[47] *Muller-Franzes G. et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis //Scientific reports. – 2023. – T. 13. – №. 1. – C. 12098.*

[48] *Huang P. et al. Chest-diffusion: A light-weight text-to-image model for report-to-cxr generation //2024 IEEE International Symposium on Biomedical Imaging (ISBI). – IEEE, 2024. – C. 1-5.*

[49] *Frid-Adar M. et al. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification //Neurocomputing. – 2018. – T. 321. – C. 321-331.*

[50] *Shin H. C. et al. Medical image synthesis for data augmentation and anonymization using generative adversarial networks //International workshop on simulation and synthesis in medical imaging. – Cham : Springer International Publishing, 2018. – C. 1-11.*

[51] *Kahn Jr C. E., Genereaux B., Langlotz C. P. Conversion of radiology reporting templates to the MRRT standard //Journal of digital imaging. – 2015. – T. 28. – №. 5. – C. 528-536.*

[52] *Hong Y. et al. Analysis of RadLex coverage and term co-occurrence in radiology reporting templates //Journal of digital imaging. – 2012. – T. 25. – №. 1. – C. 56-62.*

[53] *Peng Y. et al. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports //AMIA Summits on Translational Science Proceedings. – 2018. – T. 2018. – C. 188.*

[54] *Smit A. et al. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT //Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). – 2020. – C. 1500-1519.*

[55] *Irvin J. et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison //Proceedings of the AAAI conference on artificial intelligence. – 2019. – T. 33. – №. 01. – C. 590-597.*

[56] *Jain S. et al. Radgraph: Extracting clinical entities and relations from radiology reports //arXiv preprint arXiv:2106.14463. – 2021.*

[57] *Jang M. et al. Image Turing test and its applications on synthetic chest radiographs by using the progressive growing generative adversarial network //Scientific reports. – 2023. – T. 13. – №. 1. – C. 2356.*

[58] *Heusel M. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium //Advances in neural information processing systems. – 2017. – T. 30.*

[59] *Binkowski M. et al. Demystifying mmd gans //arXiv preprint arXiv:1801.01401. – 2018.*

[60] *Wang Z. et al. Image quality assessment: from error visibility to structural similarity //IEEE transactions on image processing. – 2004. – T. 13. – №. 4. – C. 600-612.*

[61] *Fawcett T. An introduction to ROC analysis //Pattern recognition letters. – 2006. – T. 27. – №. 8. – C. 861-874.*

[62] *Davis J., Goadrich M. The relationship between Precision-Recall and ROC curves //Proceedings of the 23rd international conference on Machine learning. – 2006. – C. 233-240.*

[63] *Shokri R. et al. Membership inference attacks against machine learning models //2017 IEEE symposium on security and privacy (SP). – IEEE, 2017. – C. 3-18.*

[64] *Hayes J. et al. Logan: Membership inference attacks against generative models //arXiv preprint arXiv:1705.07663. – 2017.*

[65] *Carlini N. et al. Extracting training data from diffusion models //32nd USENIX security symposium (USENIX Security 23). – 2023. – C. 5253-5270.*

[66] *Sajjadi M. S. M. et al. Assessing generative models via precision and recall //Advances in neural information processing systems. – 2018. – T. 31.*

[67] *Naeem M. F. et al. Reliable fidelity and diversity metrics for generative models //International conference on machine learning. – PMLR, 2020. – C. 7176-7185.*

[68] *Deo Y. et al. Metrics that matter: Evaluating image quality metrics for medical image generation //arXiv preprint arXiv:2505.07175. – 2025.*

[69] *Szegedy C. et al. Rethinking the inception architecture for computer vision //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – C. 2818-2826.*

[70] *He K. et al. Deep residual learning for image recognition //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – C. 770-778.*

[71] *Khosla P. et al. Supervised contrastive learning //Advances in neural information processing systems. – 2020. – T. 33. – C. 18661-18673.*

[72] *Efron B. Bootstrap methods: another look at the jackknife //Breakthroughs in statistics: Methodology and distribution. – New York, NY : Springer New York, 1992. – C. 569-593.*

[73] *Ibragimov A. et al. MamT 4: Multi-view Attention Networks for Mammography Cancer Classification //2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC). – IEEE, 2024. – C. 1965-1970.*