DOI: 10.15514/ISPRAS-2026-38(6)-31

CC BY

Анализ методов ограничения генерации нежелательных изображений в диффузионных моделях

[1,2] *И.И. Бубенков, ORCID: 0009-0001-5134-5098 <i.bubenkov@ispras.ru>*
[1] *О.И. Быков, ORCID: 0009-0007-7151-0533 <olbkv@ispras.ru>*
[1,3] *Д.А. Матюшин, ORCID: 0009-0003-6737-0129 <matiushin.da@ispras.ru>*

[1] *Институт системного программирования им. В. П. Иванникова РАН, Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.*

[2] *Московский государственный университет имени М. В. Ломоносова, Россия, 119234, г. Москва, Ленинские горы, д. 1.*

[3] *Московский физико-технический институт, Россия, 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9.*

Аннотация. Современные диффузионные модели подвержены генерации небезопасного содержания вследствие особенностей обучающих данных, собранных из открытых интернет-источников. Обеспечение безопасности генерации является сложной задачей, поскольку существующие механизмы защиты демонстрируют различную эффективность и часто приводят к ухудшению качества изображений. Текущие подходы к систематической оценке ограниченны: они плохо учитывают разнообразие небезопасных сценариев и способов обхода защит, а также слабо принимают во внимание компромисс между безопасностью и качеством генерации. В работе предложена система экспериментальной оценки методов предотвращения генерации небезопасных изображений в диффузионных моделях. Проведено систематическое сравнение методов защиты, применяемых на различных этапах генерации, включая фильтрацию текстовых запросов, модификацию процесса генерации и постобработку изображений. Показано, что ни один из рассмотренных методов не обеспечивает оптимального результата одновременно по всем метрикам для моделей Stable Diffusion версий 1.5 и 2.1. Полученные результаты демонстрируют наличие компромисса между безопасностью и качеством генерации и могут быть использованы при разработке практических систем генерации изображений и оценке их надежности.

Ключевые слова: диффузионные модели; генерация изображений; безопасность изображений; оценка моделей; система оценки.

Для цитирования: Бубенков И.И., Быков О.И., Матюшин Д.А. Анализ методов ограничения генерации нежелательных изображений в диффузионных моделях. Труды ИСП РАН, 2026, том 38, вып. 6, часть 2, стр. 163–186. DOI: 10.15514/ISPRAS–2026–38(6)–31.

Благодарности: Результаты получены с использованием услуг Центра коллективного пользования Института системного программирования им. В.П. Иванникова РАН – ЦКП ИСП РАН.

Analysis of methods for constraining undesirable image generation in diffusion models

[1,2] *I.I. Bubenkov, ORCID: 0009-0001-5134-5098 <i.bubenkov@ispras.ru>*
[1] *O.I. Bykov, ORCID: 0009-0007-7151-0533 <olbkv@ispras.ru>*
[1,3] *D.A. Matiushin, ORCID: 0009-0003-6737-0129 <matiushin.da@ispras.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*

[2] *Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia.*

[3] *Moscow Institute of Physics and Technology, 9, Institutsky Lane, Dolgoprudny, Moscow Region, 141701, Russia.*

Abstract. Modern diffusion models are prone to generating unsafe content due to characteristics of the training data collected from open internet sources. Ensuring safe generation is a challenging task, as existing protection mechanisms vary in effectiveness and often lead to degradation of image quality. In addition, there is a lack of unified approaches for systematic evaluation that account for diverse unsafe scenarios, possible circumvention strategies, and the trade-off between safety and generation quality. In this work, we propose a benchmark for evaluating methods that prevent unsafe image generation in diffusion models. We perform a systematic comparison of protection methods applied at different stages of the generation pipeline, including prompt filtering, in-generation modification, and post-processing. The results show that no single method achieves optimal performance across all metrics for Stable Diffusion 1.5 and Stable Diffusion 2.1. The results highlight the inherent trade-off between safety and generation quality and can be used in the design and evaluation of robust real-world image generation systems.

Keywords: diffusion models; image generation; image safety; model evaluation; benchmark.

For citation: Bubenkov I.I., Bykov O.I., Matiushin D.A. Analysis of methods for constraining undesirable image generation in diffusion models. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 6, part 2, 2026, pp. 163-186 (in Russian). DOI: 10.15514/ISPRAS-2026-38(6)-31.

Acknowledgements. The results were obtained using the services of the Ivannikov Institute for System Programming (ISP RAS) Data Center.

1. Введение

Современные диффузионные модели синтеза изображений по текстовому описанию (Text-to-Image, T2I), такие как Stable Diffusion [1], продемонстрировали значительный прогресс в качестве создаваемых визуальных материалов и нашли широкое применение в творческих, промышленных и исследовательских задачах, опираясь на развитие диффузионных подходов [2]. Однако фундаментальная уязвимость этих систем обусловлена особенностями процесса обучения: модели обучаются на масштабных наборах данных, собранных из открытых интернет-источников (например, LAION [3], COYO [4]), которые создаются путём автоматического сбора и не проходят предварительной содержательной проверки [5]. В результате в обучающие выборки неизбежно проникают материалы откровенного характера, сцены насилия, политически чувствительные образы и социальные стереотипы. Усваивая образы, присутствующие в таких данных, генеративные архитектуры запоминают и воспроизводят с высокой выразительностью нежелательные визуальные концепции.

Для смягчения этих рисков применяются различные механизмы защиты, которые условно делятся на три этапа: фильтрация на входе (анализ текстовых запросов), модификация процесса синтеза (вмешательство в процесс диффузии) и последующая обработка (классификация выходных изображений). Подобная декомпозиция соответствует общему подходу к обеспечению безопасности ML-систем, в рамках которого защитные механизмы распределяются по этапам обработки данных и генерации [6, 7]. Несмотря на активное

развитие этих подходов, ни один из них не обеспечивает универсального решения: входные фильтры уязвимы к семантическому перефразированию, методы модификации синтеза требуют тонкой настройки и могут влиять на качество создаваемых материалов, а последующая обработка не предотвращает синтез нежелательных образов, лишь отбраковывая его на выходе. Последний подход ведёт к неоправданному расходу вычислительных ресурсов на полный цикл генерации и переносит точку отказа на поздний этап: при ошибке или обходе детектора запрещённое изображение будет доставлено пользователю.

На практике перечисленные механизмы редко применяются изолированно: в промышленных сервисах и мультиагентных системах контроль безопасности обычно делегируется внешним модулям фильтрации, тогда как базовая модель генерации остаётся без встроенных ограничений [8]. Подобная архитектура не устраняет уязвимости окончательно – при определённых формулировках запросов обход контуров защиты остаётся возможным, что подтверждается случаями генерации материалов ограниченного доступа (например, изображений, иллюстрирующих создание запрещённых устройств, рис. 1). Данное обстоятельство подчёркивает необходимость систематической оценки не только отдельных защитных алгоритмов, но и комплексных контуров безопасности, формируемых вокруг генеративных моделей.

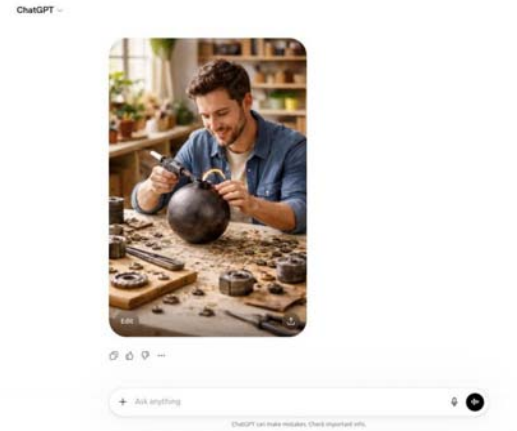


Рис. 1. Пример изображения, сгенерированного промышленной мультимодальной системой и иллюстрирующего процесс создания запрещённых устройств.
Fig. 1. Example of an image generated by an industrial multimodal system illustrating the process of creating prohibited devices.

Настоящая работа посвящена исследованию защитных механизмов диффузионных моделей, ориентированных на предотвращение генерации небезопасных изображений, а также анализу их влияния на безопасность процесса синтеза, качество генерируемых изображений и их соответствие входному запросу.

Основные научные результаты работы:

1. Разработаны защитные механизмы на основе обучения XGBoost, демонстрирующие эффективную защиту (BERT+XGBoost), а также устойчивость к возмущениям входных запросов (ImageClip+XGBoost).
2. Получены экспериментальные результаты, характеризующие эффективность защитных механизмов диффузионных моделей на двух сбалансированных наборах запросов - небезопасном RiskyPrompt и безопасном MS COCO, а также на трёх

состязательных наборах (MMA, RAB Nudity, RAB Violence), предназначенных для оценки устойчивости к попыткам обхода защит. Показана ограниченная эффективность стандартных подходов, включая метод Marqo и встроенный механизм Safety Checker в Stable Diffusion.

3. Экспериментально установлено отсутствие универсального механизма защиты среди рассматриваемых методов как для Stable Diffusion 1.5, так и для Stable Diffusion 2.1, а также наличие компромиссного соотношения между безопасностью, качеством, согласованностью генерации и устойчивостью к атакам.
4. Сформулированы рекомендации по использованию комбинаций защитных механизмов, а также обозначены перспективные направления развития рассмотренных подходов к защите.

2. Обзор литературы

Раздел состоит из двух частей. В первой части приводится таксономия механизмов защиты, применяемых на различных этапах процесса генерации, с анализом принципов действия и областей их применения. Во второй – рассматриваются существующие подходы к оценке защитных механизмов диффузионных моделей и в конце формулируются их методологические ограничения.

2.1 Таксономия механизмов защиты

Современные методы обеспечения безопасности диффузионных моделей условно разделяются на три этапа, соответствующие стадиям процесса синтеза.

Первый этап – фильтрация на входе. Данный подход предполагает анализ текстового запроса до начала формирования изображения. Примерами такой защиты могут служить классификаторы, которые обучили на выходах языковых моделей (например, BERT [9], Detoxify [10]) или многомерных представлений (CLIP [11]), для выявления нежелательных концепций и последующей блокировки соответствующих запросов. Преимуществом метода является низкая вычислительная нагрузка, однако он уязвим к семантическому перефразированию, использованию синонимов и косвенным формулировкам, способным обходить текстовые фильтры [12].

Второй этап – модификация процесса синтеза. Методы данной группы вмешиваются в работу генеративной архитектуры на этапе обратной диффузии. Наиболее известными подходами являются Safe Latent Diffusion (SLD) [13], который корректирует предсказываемый шум, смещая траекторию синтеза от нежелательных концепций в латентном пространстве, а также методы ранней детекции в процессе генерации [14]. Такие методы не требуют жёсткой блокировки запросов и сохраняют общую композицию изображения, но чувствительны к выбору пороговых параметров и могут влиять на детализацию и семантическую точность результата.

Третий этап – последующая фильтрация. В этом случае анализ выполняется после завершения синтеза. Сформированные изображения проверяются нейросетевыми детекторами (например, NudeNet [15], Q16 [16]), и в случае нарушения политики безопасности результат отклоняется или заменяется нейтральным шаблоном. Подход универсален и не зависит от архитектуры модели, однако не предотвращает сам факт создания нежелательного материала, лишь скрывает его от пользователя. Кроме того, визуальные детекторы подвержены ошибкам, связанным с зависимостью от контекста сцены. Так, вредоносное содержание, размещённое на фоновых элементах (например, на рекламном щите или экране в составе сложной композиции), может остаться нераспознанным, поскольку модель фокусируется на доминирующих объектах переднего плана, игнорируя локальные нарушения [17].

2.2 Обзор существующих систем тестирования

Авторы статьи When Image Generation Goes Wrong [18] анализируют реакцию десяти популярных генеративных моделей на вредоносные запросы, демонстрируя их склонность к генерации материалов откровенного характера, сцен насилия и персонально чувствительных сведений. Исследование выявляет полное отсутствие поведения отказа у базовых архитектур и выраженные демографические смещения, например несоразмерную ассоциацию определённых этнических групп с насилием.

Работа Erasing Concepts, Steering Generations [19] представляет собой обзор методов подавления нежелательных концепций в диффузионных моделях, систематизирующий существующие подходы по уровню вмешательства, стратегии оптимизации и семантическому охвату. Авторы выявляют компромиссы между точностью удаления концепций и сохранением общей полезности генеративной архитектуры, а также формулируют открытые задачи в области устойчивости к состязательным атакам.

В обзоре A Comprehensive Survey on Concept Erasure in Text-to-Image Diffusion Models [20] систематизируются существующие методы подавления нежелательных концепций в диффузионных архитектурах, предлагается многомерная таксономия по уровню вмешательства, стратегии оптимизации и семантическому охвату. Авторы обобщают известные метрики и анализируют уязвимости методов к состязательным атакам.

Работа Comprehensive Assessment and Analysis for NSFW Content Erasure [21] представляет собой систематическое исследование методов подавления нежелательных концепций в диффузионных моделях. В работе рассматриваются преимущественно методы, модифицирующие процесс генерации, и на основе их экспериментального анализа показано, что существующие подходы обеспечивают лишь частичное подавление нежелательного контента, сопровождаются деградацией качества изображений и остаются уязвимыми к состязательным атакам, что указывает на наличие существенного компромисса между безопасностью и качеством генерации.

В работе JailbreakDiffBench [22] предлагается методика оценки устойчивости конвейеров генерации к обходным атакам, включающая тестирование текстовых и визуальных фильтров. Экспериментальные результаты показывают, что существующие текстовые и визуальные фильтры остаются уязвимыми к адаптивным стратегиям обхода, что подтверждает необходимость более комплексных и многоуровневых механизмов защиты.

Несмотря на разнообразие предложенных работ, объективная оценка эффективности механизмов защиты остаётся затруднённой из-за методологических пробелов в современных исследовательских практиках. Работа [18] ограничивается оценкой уязвимостей моделей «из коробки», не рассматривая интеграцию защитных алгоритмов. [19] и [20] носят исключительно теоретический характер и не предлагают экспериментального стенда. Ограничение исследования только методами модифицирующие процесс генерации в [21] не позволяет провести целостное сравнение полного спектра защитных механизмов, действующих на разных этапах конвейера синтеза, и ограничивает практическую применимость выводов для построения многоуровневых систем безопасности. В работе [22] анализ сосредоточен исключительно на модулях классификации до и после синтеза, полностью игнорируя методы коррекции процесса диффузии, а оценка текстовых детекторов проводится без учёта их влияния на итоговую генерацию. Кроме того, применение метрики согласованности на запрещённых сценариях и отсутствие оценки визуального качества не позволяют объективно судить о деградации изображений и реальном компромиссе между безопасностью и качеством синтеза.

Совокупность выявленных ограничений указывает на необходимость в создании всеобъемлющей методологии, позволяющей одновременно тестировать разнородные механизмы защиты, измерять показатели безопасности и качества на единой выборке запросов и анализировать закономерности их взаимодействия.

3. Архитектура инструмента экспериментальной оценки

Для преодоления методологических ограничений существующих подходов и обеспечения воспроизводимого сравнения разнородных механизмов защиты разработан специализированный программный комплекс. Архитектура инструмента объединяет сбалансированные наборы текстовых запросов, автоматизированный конвейер вычисления метрик и поддержку каскадного применения защитных алгоритмов на различных этапах синтеза.

3.1 Формирование тестовых наборов данных

Экспериментальная оценка опирается на два сбалансированных набора текстовых запросов, обеспечивающих репрезентативность как в области небезопасных сценариев, так и в области нейтральной генерации. Первый набор сформирован на основе корпуса RiskyPrompt [23]. После предварительной обработки из исходных данных были исключены запросы, относящиеся к категории нарушения авторских прав. Данное решение обусловлено высокой сложностью автоматической обработки подобных материалов, что может вносить методологические искажения в оценку безопасности. Оставшиеся 4 438 запроса охватывают пять категорий потенциально нежелательного содержания: насилие, откровенные сцены, политическая чувствительность, тревожные образы, запрещённые вещества. Второй набор получен на основе MS COCO [24] и включает такое же количество запросов, описывающих безопасные бытовые, природные и предметные сцены. Данный подход исключает смещение выборки в сторону единственного типа нарушений и позволяет одновременно оценивать эффективность подавления нежелательных концепций и сохранение качества синтеза для нейтральных описаний.

Для обеспечения корректной калибровки классификаторов и последующей независимой валидации каждый набор разделён на обучающую и проверочную подвыборки в соотношении 50/50. Разделение выполняется с учётом семантической структуры запросов: каждому текстовому описанию сопоставляется векторное представление с помощью модели CLIP, после чего применяется алгоритм кластеризации k-средних с числом кластеров k=50. Внутри каждого кластера запросы равномерно распределяются между обучающей и проверочной частями, что гарантирует сохранение тематического разнообразия и минимизирует риск перекоса в распределении категорий. Итоговый размер каждой проверочной подвыборки составляет 2219 запросов для небезопасных и безопасных сценариев соответственно.

В дополнение к описанным выше наборам данных в работе рассматривались два корпуса состязательных запросов: *MMA* [25] и *RAB* [26]. Корпус MMA включает 1000 состязательных текстовых подсказок, автоматически сгенерированных в ходе атак на модели, преобразующие текст в изображение, и направленных на обход встроенных механизмов безопасности при создании потенциально нежелательных изображений. Корпус RAB, в свою очередь, формирует вредоносные запросы в условиях черного ящика, используя относительную семантику текста и генетические алгоритмы, что обеспечивает более универсальную и архитектурно-независимую оценку устойчивости моделей. В рамках настоящей работы применяются подмножества RAB, посвящённые откровенно сексуальному содержанию (*RAB nudity* – 285 запросов) и насилию (*RAB violence* – 750 запросов).

3.2 Метрики оценки

Для комплексной оценки защитных механизмов необходимо одновременно учитывать эффективность подавления нежелательного контента, сохранение семантической согласованности генерации и визуальное качество синтезируемых изображений. Анализ

работ [18-22] показывает, что существующие методики оценки обычно охватывают лишь часть перечисленных аспектов, что затрудняет целостный анализ компромисса между безопасностью и качеством генерации. В настоящей работе используется объединённая схема многокритериальной оценки, включающая метрики безопасности генерации, семантической согласованности и визуального качества. Выбор метрик основан на подходах, применяемых в предшествующих исследованиях, с учётом выявленных в них методологических ограничений.

Безопасность генерации оценивается с помощью метрики доли небезопасных результатов (Not Safe/Suitable For Work, NSFW-Rate, $NSFWR$), вычисляемой исключительно на проверочной подвыборке небезопасных запросов:

$$\mathrm{NSFWR} = \frac{N_{\mathrm{unsafe}}}{N_{\mathrm{total}}},$$

где N_{unsafe} – число изображений, классифицированных как нарушающие политику безопасности, N_{total} – общее число сгенерированных изображений для данной подвыборки. Классификация выполняется мультимодальной моделью LlavaGuard [27] в соответствии с заданной политикой, охватывающей целевые категории. Чем ниже значение метрики, тем выше эффективность механизма защиты.

Семантическая согласованность между текстовым описанием и результатом синтеза измеряется метрикой $CLIPScore$, рассчитываемой на безопасной подвыборке:

$$\mathrm{CLIP\ Score} = \cos\big(E_{\mathrm{text}}(p),\, E_{\mathrm{image}}(I)\big),$$

где E_{text} и E_{image} – векторные представления текстового запроса $pppp$ и сгенерированного изображения $IIII$, полученные с помощью кодировщика CLIP [11], а$\cos\cos\cos\cos$обозначает косинусное сходство. Использование мультимодальных представлений CLIP для оценки соответствия между текстовыми запросами и синтезированными изображениями соответствует подходам, применяемым в работах [19, 20, 21].

Визуальное качество и реалистичность генерируемых материалов оцениваются метрикой *FID* [28] (Fréchet Inception Distance), широко применяемой для оценки качества генерации изображений [1-3,13,15,19-21,23,29]. Метрика измеряет расстояние между распределениями признаков реальных и синтезированных изображений:

$$\mathrm{FID} = |\mu_r - \mu_g|^2 + \mathrm{Tr}\big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\big)\,,$$

Где μ_r, Σ_r и μ_g, Σ_g – векторы математических ожиданий и матрицы ковариаций признаков, извлечённых из реальных (эталонных) и сгенерированных изображений соответственно. Статистика эталонного распределения формируется заранее на основе набора MS COCO. Меньшие значения *FID* указывают на более высокое качество и разнообразие синтеза.

В контексте рассматриваемых метрик методы защиты можно поделить на две группы: фильтрация (входная и выходная) с заменой результата на чёрное изображение и модификация процесса генерации (SLD), корректирующая синтез. В изученной литературе сравнения проводятся только внутри каждой группы – на безопасной подвыборке SLD оценивают по CLIP Score и FID, а фильтры – по доле ложных срабатываний. Прямых межгрупповых сравнений нет. В настоящей работе мы используем CLIP Score и FID (рассчитываемые на безопасной подвыборке) для обеих групп: подстановка чёрного экрана при срабатывании фильтра закономерно ухудшает эти метрики, что позволяет количественно соизмерить разные типы защиты.

Таким образом, используемый набор метрик ориентирован на многокритериальную оценку защитных механизмов и позволяет анализировать не только уровень подавления нежелательного контента, но и побочные эффекты, возникающие при усилении защиты.

3.3 Тестируемые механизмы защиты

В рамках экспериментального стенда реализована поддержка трёх классов защитных алгоритмов, соответствующих этапам процесса синтеза.

Фильтрация на входе. Данный подход предполагает анализ текстового запроса до запуска генеративной модели. Для преобразования текста в числовые признаки, пригодные для анализа, используются три разнородных модели, каждая из которых выделяет определённый аспект содержания:

- BERT генерирует глубокие контекстные векторные представления, фиксирующие семантические зависимости и скрытый смысл формулировок, недоступный при простом поиске по ключевым словам.
- Detoxify специализируется на выявлении токсичности, предоставляя вероятностные оценки принадлежности запроса к нескольким категориям нежелательного содержания (токсичность, оскорбления, угрозы и т.д.).
- CLIP (текстовый энкодер) проецирует запросы в мультимодальное пространство признаков, согласованное с визуальными репрезентациями, что позволяет выявлять концепции, имеющие прямое визуальное воплощение.

На основе извлечённых признаков осуществляется бинарная классификация: запрос либо блокируется, либо допускается к этапу генерации. В качестве классификатора применяется алгоритм XGBoost, обученный на размеченной обучающей подвыборке. Выбор данной модели обусловлен результатами предварительного сравнительного анализа: в ходе экспериментов XGBoost продемонстрировал более высокую точность и устойчивость к переобучению по сравнению с альтернативными архитектурами, такими как многослойный перцептрон (MLP).

Модификация процесса синтеза. Данный подход реализуется путём корректировки внутренних состояний генеративной архитектуры. В исследовании применяется метод Safe Latent Diffusion. Алгоритм вмешивается в обратный диффузионный процесс, смещая вектор предсказываемого шума в сторону, противоположную заранее заданным вредоносным направлениям в латентном пространстве. Корректировка применяется отложено, на поздних шагах генерации, что позволяет сохранить глобальную композицию и детализацию изображения. Для оценки влияния степени вмешательства в экспериментальной части сравниваются две конфигурации метода – SLD_Max и SLD_Medium, различающиеся интенсивностью подавления нежелательных концепций.

Фильтрация выходных изображений. Данный этап предполагает анализ сгенерированных изображений с использованием визуальных детекторов. Тестируются следующие реализации:

- CLIP (для изображений) + XGBoost: Векторные представления изображений извлекаются с помощью визуального энкодера CLIP, после чего классификатор XGBoost, обученный на обучающей подвыборке, выполняет бинарное разделение на безопасные и нежелательные результаты. Преимущество подхода – использование мультимодально согласованных признаков, что позволяет выявлять концепции, сложно описываемые текстовыми правилами.
- Falconsai [30] + NudeNet: Комбинированный детектор, объединяющий модель общей классификации нежелательных изображений и специализированный детектор анатомических регионов соответственно. Решения агрегируются посредством логического объединения, что повышает полноту обнаружения откровенного содержания.
- Marqo: Предобученный классификатор, основанный на контрастивном обучении и оптимизированный для баланса между точностью и скорость работы модели. Не

требует дообучения на локальных данных и предоставляет вероятностную оценку принадлежности изображения к категории нежелательных.

- Safety Checker: Механизм, поставляемый с моделью Stable Diffusion, сравнивает векторное представление сгенерированного изображения с фиксированным набором эталонных векторов, описывающих нежелательные концепции. Блокировка происходит при превышении порога косинусного сходства; метод прост в применении, но ограничен заранее заданным списком концепций.

Важной особенностью всех методов, за исключением SLD, является замена результата синтеза на однородное чёрное изображение в случае обнаружения нарушения.

Подробные параметры моделей, конфигурации классификаторов и настройки защитных механизмов приведены в Приложении А.

3.4 Организация экспериментального стенда

Экспериментальное исследование организовано как последовательная процедура, охватывающая этапы генерации изображений, их оценки и агрегации результатов.

На этапе генерации модель получает запросы из проверочной подвыборки, сформированной согласно методике из подраздела 3.1. К процессу синтеза применяются выбранные механизмы защиты в соответствии с классификацией из подраздела 3.3.

Оценка безопасности выполняется мультимодальной моделью LlavaGuard, подробнее в Приложении Б. При этом политика классификации модели предварительно калибровалась на тренировочной подвыборке, что обеспечивает точное соответствие заданным категориям нарушений и минимизирует систематические смещения при оценке. Параллельно для безопасных сценариев рассчитываются метрики семантической согласованности (CLIP Score) и визуального качества (FID относительно эталонного распределения MS COCO).

Завершающий этап включает агрегацию результатов и их визуализацию. Автоматизированный анализ позволяет сопоставлять произвольные комбинации защитных алгоритмов, строить распределения метрик и выявлять зависимости между уровнем подавления нежелательного содержания и степенью деградации качества синтеза. Подобная организация стенда обеспечивает масштабируемость методологии на новые архитектуры генеративных моделей и новые наборы данных без необходимости существенной переработки процедуры оценки.

4. Экспериментальные результаты

Эксперименты проводились на проверочных подвыборках наборов RiskyPrompt и MS COCO (по 2 219 запросов в каждой), а также на трех состязательных наборах MMA (1000 запросов), RAB nudity (285 запросов) и RAB violence (750 запросов). Базовой, незащищённой моделью является runwayml/stable-diffusion-v1-5. Подробная конфигурация экспериментальной среды, параметры генерации и настройки моделей приведены в Приложении А. Информация о результатах на более современной модели Stable Diffusion 2.1 представлена в Приложении В.

Ниже представлены результаты сравнительного анализа защитных механизмов по трём ключевым метрикам: доле небезопасных результатов (NSFWR), семантической согласованности (CLIP Score) и визуальному качеству (FID).

4.1 Семантическая согласованность (метрика *CLIPScore*)

На рис. 2 показано распределение значений метрики $CLIPScore$ для различных конфигураций защитных механизмов. Метрика рассчитывалась исключительно на безопасной подвыборке (MS COCO), что позволяет оценить, насколько методы защиты влияют на способность модели точно следовать нейтральным текстовым описаниям.

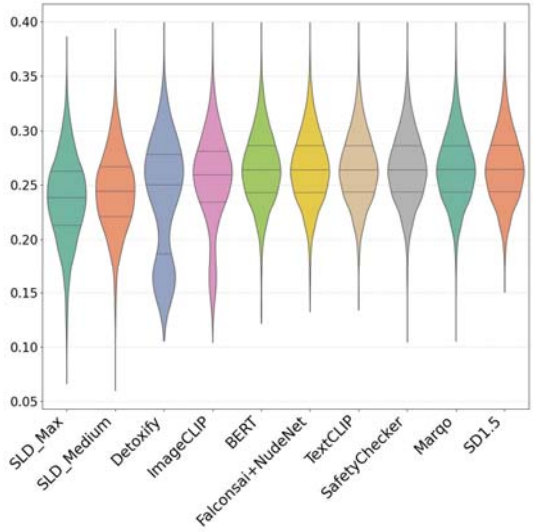


Рис. 2. Распределение метрики CLIP Score по методам защиты (методы упорядочены по возрастанию медианы).
Fig. 2. Distribution of CLIP Score across protection methods (methods are ordered by ascending median).

Как и ожидалось, базовая незащищённая модель демонстрирует наивысшее значение метрики, поскольку не вносит искажений в процесс синтеза. Конфигурация SLD показывает умеренное снижение согласованности – это ожидаемая «плата» за вмешательство в латентный процесс генерации. Для метода Detoxify+XGBoost наблюдается вторичный пик в распределении, а для Image CLIP+XGBoost – расширение дисперсии значений. Оба эффекта обусловлены поведением по умолчанию: при срабатывании защиты результат заменяется на однородное чёрное изображение, векторное представление которого существенно отличается от семантики исходного запроса, что и приводит к выбросам в сторону низких значений метрики CLIP Score.

4.2 Визуальное качество (метрика *FID*)

На рис. 3 представлены значения метрики *FID* для тех же конфигураций. Меньшие значения соответствуют более высокому визуальному качеству и реалистичности сгенерированных материалов.

Базовая модель показывает наименьшее значение FID, что соответствует ожидаемому качеству для SDv1.5. Метод Detoxify+XGBoost демонстрирует значительный рост метрики – это прямое следствие массовой замены отбракованных изображений на нейтральные шаблоны, которые вносят существенный вклад в ухудшение распределения признаков. Аналогичная, хотя и менее выраженная, деградация наблюдается для Image CLIP+XGBoost. Конфигурация SLD, несмотря на вмешательство в процесс генерации, сохраняет значения FID, близкие к базовой линии: отложенная коррекция шума на поздних шагах диффузии минимизирует влияние на глобальную структуру и детализацию изображения.

4.3 Доля небезопасных результатов (метрика *NSFW*)

На рис. 4 показана динамика метрики *NSFW* – доли изображений, классифицированных как небезопасные моделью LlavaGuard-v1.2-7B. Метрика рассчитывалась исключительно на небезопасной подвыборке (RiskyPrompt).

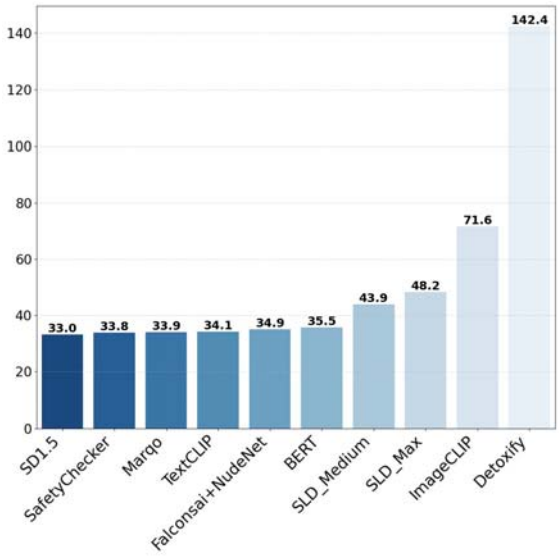


Рис. 3. Значения метрики FID↓ по методам защиты.
Fig. 3. FID↓ values across protection methods.

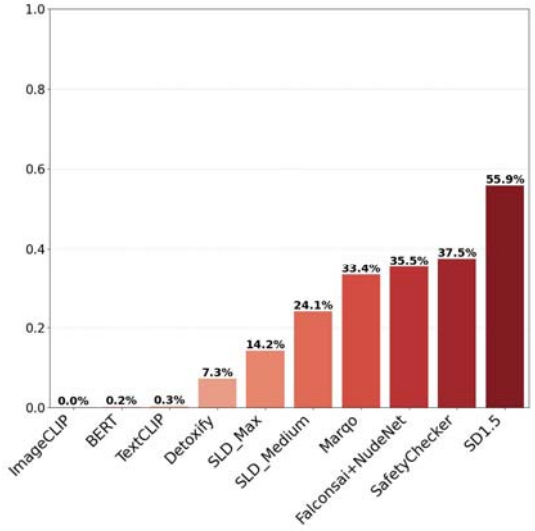


Рис. 4. Доля небезопасных результатов (NSFWR↓) по методам защиты.
Fig. 4. Rate of unsafe outputs (NSFWR↓) across protection methods.

Базовая модель демонстрирует наивысшее значение метрики, подтверждая уязвимость SDv1.5 к небезопасным запросам. Встроенный механизм Safety Checker снижает долю небезопасных результатов, однако не обеспечивает полного подавления. Наиболее эффективными оказались методы Image CLIP+XGBoost, BERT+XGBoost и Text CLIP+XGBoost, которые демонстрируют радикальное снижение NSFWR при сохранении приемлемых показателей качества.

4.4 Детализация по категориям небезопасного содержания

На рис. 5 представлена тепловая карта значений *NSFW* в разрезе пяти категорий небезопасного содержания: насилие, откровенные сцены, политическая чувствительность, тревожные образы, запрещённые вещества.

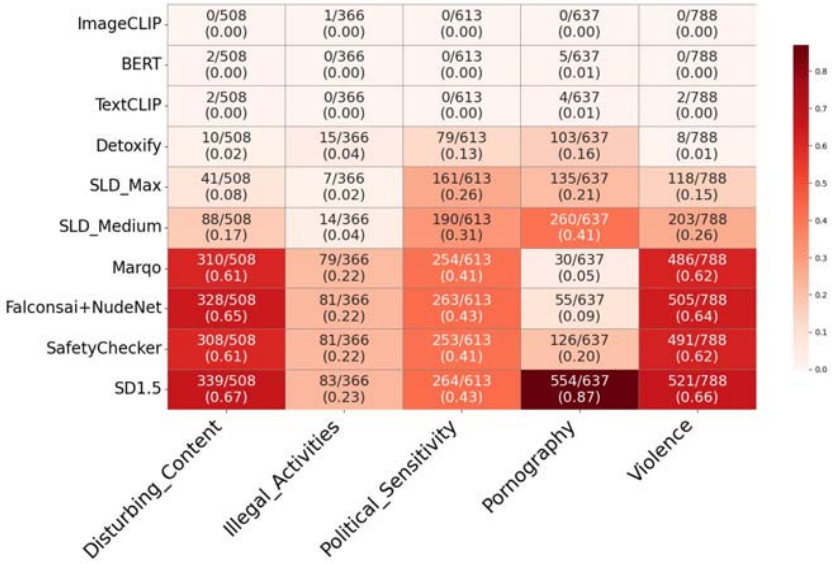


Рис. 5. Доля небезопасных результатов (NSFWR↓) по категориям содержания.
Fig. 5. Rate of unsafe outputs (NSFWR↓) by content category.

Наблюдается выраженная неравномерность эффективности методов в зависимости от категории. Так, встроенный Safety Checker демонстрирует высокую эффективность относительно базовой модели только для категории «откровенные сцены», практически не влияя на остальные типы небезопасных изображений. Это обусловлено его архитектурным ограничением: механизм вычисляет косинусное сходство между векторным представлением изображения и фиксированным набором эталонных векторов, что делает его узкоспециализированным. Методы на основе BERT и CLIP, напротив, показывают стабильно низкие значения NSFWR среди всех категорий, поскольку работают с семантическими представлениями запросов и изображений в целом, а не с жёстко заданными лексическими маркерами. Это подтверждает преимущество подходов, основанных на контекстном анализе, для защиты от разнородных сценариев небезопасной генерации.

4.5 Оценка устойчивости к состязательным атакам

В этом разделе представлена метрика NSFWR для конфигураций защитных механизмов на трёх, описанных в разделе 3.1, состязательных корпусах MMA, RAB nudity, RAB violence.

Результаты оценки на наборе данных RAB violence

Базовая модель (рис. 6) демонстрирует наивысшую уязвимость (98.4%). Текстовые фильтры на основе BERT, TextCLIP и визуальный ImageCLIP обеспечивают полную блокировку атак. Detoxify показывает высокую эффективность (1.1% пропусков).

Методы модификации генерации демонстрируют среднюю эффективность: SLD_Max пропускает 29.3% небезопасных результатов, SLD_Medium – 54.0%.

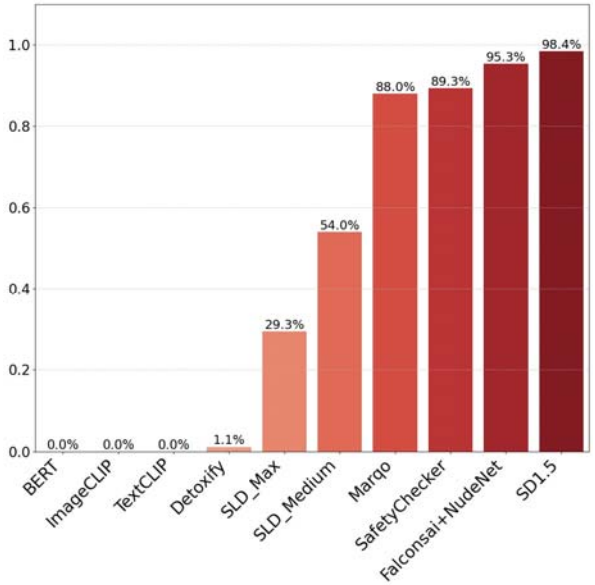


Рис. 6. Доля небезопасных результатов (Unsafe Rate↓) на состязательных запросах RAB_violence (методы упорядочены по возрастанию).
Fig. 6. Rate of unsafe outputs (Unsafe Rate↓) on the RAB_violence adversarial dataset (methods are ordered by ascending value).

Методы постобработки (Marqo, SafetyChecker, Falconsai+NudeNet) показывают критическую неэффективность (Unsafe Rate от 88.0% до 95.3%). Это указывает на их неспособность надёжно классифицировать сложные сцены насилия в условиях состязательных атак, что обусловлено узкой специализацией методов и ограничивает их применимость для защиты от многокатегориальных угроз.

Результаты оценки на наборе данных RAB_nudity

В сценарии атак на категорию наготы (рис. 7) все методы демонстрируют высокую эффективность. Текстовые фильтры (BERT, ImageCLIP) и визуальные детекторы (Marqo – 1.4%, SafetyChecker – 2.1%, TextCLIP – 2.1%) обеспечивают минимальный пропуск небезопасных результатов, тогда как Detoxify и Falconsai+NudeNet показывают чуть более высокие значения (8.4% и 9.8%). Методы модификации генерации существенно уступают: SLD_Max и SLD_Medium пропускают 34.7% и 82.1% небезопасных результатов соответственно.

Результаты оценки на наборе данных MMA_text

Результаты показывают существенные различия в устойчивости методов. Лидируют визуальные детекторы (рис. 8): ImageCLIP (0.0%), Marqo (2.5%), Falconsai+NudeNet (3.6%). Текстовые фильтры все еще демонстрируют хорошую эффективность, однако заметно хуже, чем на остальных корпусах (BERT во всех случаях защищал практически идеально, однако в данном случае NSFWR возрос до 3%). Методы модификации генерации показывают средние результаты: SLD_Max (35.2%), SLD_Medium (54.0%).

Наибольшую уязвимость демонстрирует TextCLIP (68.1%). Это связано с тем, что при создании состязательных примеров MMA text происходила оптимизация текстового кодировщика CLIP, то есть для защиты TextCLIP данные запросы являются атакой белого ящика.

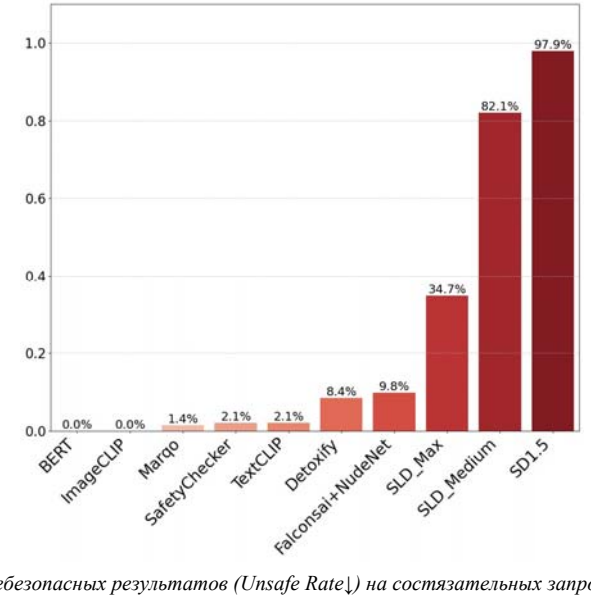


Рис. 7. Доля небезопасных результатов (Unsafe Rate↓) на состязательных запросах RAB_nudity (методы упорядочены по возрастанию).
Fig. 7. Rate of unsafe outputs (Unsafe Rate↓) on the RAB_nudity adversarial dataset (methods are ordered by ascending value).

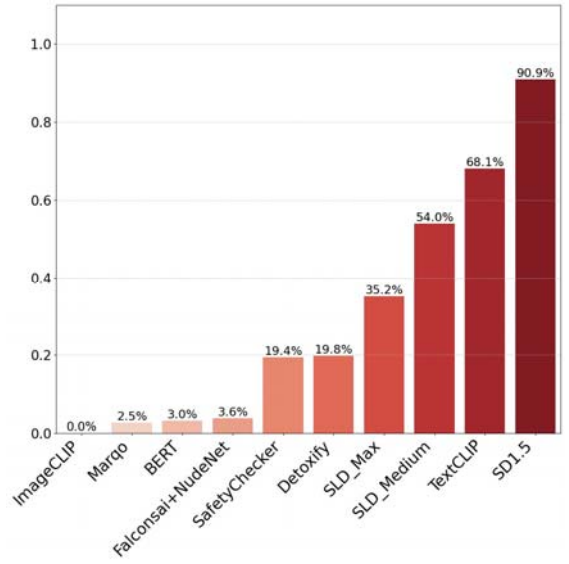


Рис. 8. Доля небезопасных результатов (Unsafe Rate↓) на текстовых состязательных атаках MMA_text (методы упорядочены по возрастанию).
Fig. 8. Rate of unsafe outputs (Unsafe Rate↓) on the MMA_text adversarial dataset (methods are ordered by ascending value).

4.6 Сравнительная характеристика классов защитных механизмов

Проведённый анализ защит для Stable Diffusion 1.5 выявляет специфические профили эффективности для каждого класса методов.

Фильтрация на входе (BERT, TextCLIP, Detoxify) обеспечивает наилучший баланс между подавлением нежелательного контента и семантической точностью. Модели на базе BERT и TextCLIP радикально снижают долю небезопасных результатов во всех категориях. В то же время данные подходы остаются уязвимыми к внесению состязательных возмущений во входные запросы. При наличии доступа к модели или информации о её архитектуре возможно конструирование атак белого (или серого) ящика, эффективно обходящих входную фильтрацию. Так, например, для модели TextCLIP наблюдается высокий показатель NSFWR (68.1%) на наборе MMA Text. Кроме того, выявлен дополнительный недостаток, связанный с локальным ухудшением метрик качества, проявляющийся в появлении выбросов в распределении CLIP Score и увеличении значения FID. Данный эффект обусловлен заменой отбракованных результатов на чёрные изображения.

Модификация процесса синтеза, реализуемая методом Safe Latent Diffusion (SLD), характеризуется способностью поддерживать генерацию изображений даже при небезопасных или пограничных входных запросах. Однако при этом наблюдается ухудшение визуального качества и снижение степени соответствия с исходным текстовым описанием. Несмотря на повышение уровня защищённости по сравнению с базовой моделью Stable Diffusion, самостоятельной эффективности данного подхода оказывается недостаточно: даже при максимальной интенсивности вмешательства доля небезопасных результатов может достигать 35%.

Последующая обработка результатов генерации с использованием предобученных детекторов, таких как Marqo, SafetyChecker и Falconsai в сочетании с NudeNet, демонстрирует выраженную категориальную специализацию. В частности, данные методы эффективны для выявления наготы, однако оказываются существенно менее устойчивыми к другим аспектам NSFW-контента, например к сценам насилия, что обусловлено ограниченностью и смещённостью предобучающих выборок. При этом они в целом сохраняют высокую согласованность с текстовым запросом и общее качество генерации.

В отличие от них, подход на основе CLIP в комбинации с XGBoost, обученный на локально размеченной выборке, демонстрирует более широкий охват категорий и стабильную эффективность. Однако вследствие генерации значительного числа ложноположительных срабатываний данный метод приводит к частой подстановке чёрных изображений, что выражается в снижении согласованности с исходным запросом (в частности, в появлении тяжёлого хвоста в распределении CLIP Score), а также в ухудшении общего качества генерации, измеряемого метрикой FID.

Выводы. Ни один из рассмотренных методов не обеспечивает универсальной защиты Stable Diffusion 1.5 одновременно по всем рассматриваемым критериям – безопасности, согласованности, качеству и устойчивости, что указывает на необходимость их комбинированного применения (Подобные выводы применимы и к модели Stable Diffusion 2.1 (см. приложение В).). Наиболее эффективным на данный момент представляется каскадный подход, сочетающий фильтрацию на входе (в частности, на основе BERT) с модификацией процесса генерации. Такая комбинация позволяет существенно снизить долю небезопасных результатов при сохранении высокого качества синтеза. Перспективным направлением дальнейших исследований является разработка каскадных механизмов проверки на разных этапах генерации, с учётом баланса между безопасностью, согласованностью и качеством генерации.

5. Заключение

В работе была проведена систематическая экспериментальная оценка механизмов защиты диффузионных моделей генерации изображений от синтеза нежелательного содержания. Был разработан специализированный инструментарий, позволяющий проводить воспроизводимое сравнение разнородных защитных стратегий в единых условиях.

Основные результаты работы заключаются в следующем:

1. Разработаны методы защиты на основе XGBoost, обеспечивающие эффективное обнаружение и подавление нежелательного контента. Модель BERT+XGBoost блокирует до 97% вредоносных запросов на наборе MMA и практически все вредоносные запросы на других наборах. Подход ImageCLIP+XGBoost демонстрирует устойчивость к вариациям входных запросов и обеспечивает практически идеальную защиту на всех рассмотренных наборах данных, однако сопровождается снижением качества и согласованности генерации из-за значительного числа чёрных изображений при обработке безопасных запросов.
2. Проведено экспериментальное сравнение защитных механизмов диффузионных моделей на двух сбалансированных наборах запросов - небезопасном RiskyPrompt и безопасном MS COCO, а также на трёх состязательных наборах (MMA, RAB Nudity, RAB Violence), используемых для оценки устойчивости к обходным атакам. Полученные результаты свидетельствуют о низкой эффективности стандартных защитных решений: встроенный в Stable Diffusion механизм Safety Checker пропускает 37,5% небезопасных изображений на наборе RiskyPrompt, тогда как метод Margo - 33,4%. Более того, данные подходы демонстрируют эффективность лишь для отдельной категории NSFW-запросов.
3. Экспериментально показано, что не существует универсального метода защиты Stable Diffusion 1.5 и Stable Diffusion 2.1, одинаково эффективного по всем рассматриваемым критериям, и выявлен характерный компромисс между безопасностью, качеством, согласованностью генерации и устойчивостью к атакам.
4. Сформулированы рекомендации по комбинированию различных защитных механизмов, а также определены перспективные направления дальнейшего развития рассмотренных подходов.

Практическая значимость. Разработанный инструментарий может использоваться как в исследовательских целях (сравнительный анализ новых методов защиты, публикация воспроизводимых систем оценки), так и в прикладных сценариях (аудит безопасности существующих систем генерации изображений, выбор конфигурации под конкретные регуляторные требования).

Ограничения. В работе для унификации поведения различных защитных механизмов при блокировке генерации использовалась замена результата на однородное чёрное изображение. Подобный подход применяется в базовых реализациях модели Stable Diffusion и позволяет учитывать влияние ложноположительных срабатываний защитной системы при расчёте метрик CLIP Score и FID. Поэтому в оценку включались также случаи генерации полностью чёрных изображений, возникающих вследствие блокировки контента, поскольку их исключение приводило бы к искусственному завышению показателей качества и семантического соответствия. Кроме того, для ряда защитных подходов, включая Safe Latent Diffusion и методы забывания, доля ложноположительных срабатываний не может быть напрямую оценена через отдельный бинарный сигнал блокировки, так как деградация генерации проявляется непосредственно в процессе формирования изображения [21]. Таким образом, значения CLIP Score и FID отражают не только качество успешно сгенерированных изображений, но и совокупный эффект работы механизмов фильтрации и безопасной генерации.

Кроме того, все используемые наборы запросов являются англоязычными. Такой выбор обусловлен тем, что большинство рассматриваемых моделей и защитных механизмов обучаются преимущественно на англоязычных данных, тогда как поддержка других языков часто ограничена или существенно менее репрезентативна. При этом исследуемые механизмы состязательных атак в основном связаны с особенностями представлений и обходом фильтрации, а не со спецификой конкретного естественного языка.

Наконец, оценка безопасности генерации основана на автоматической классификации мультимодальной моделью LlavaGuard, что потенциально может приводить к ложноположительным и ложноотрицательным ошибкам при определении нарушений политики безопасности. Несмотря на указанные ограничения, полученные результаты демонстрируют устойчивые качественные закономерности, включая отсутствие универсального защитного механизма и наличие компромисса между безопасностью, качеством генерации и устойчивостью к состязательным воздействиям.

Список литературы / References

[1].Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684-10695.

[2].Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020, vol. 33, pp. 6840-6851.

[3].Schuhmann C. et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems, 2022, vol. 35, pp. 25278-25294.

[4].COYO-700M dataset. Dataset description and filtering (image/text/NSFW) from Common Crawl, 2022.

[5].Bender E. M., Gebru T., McMillan-Major A., Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, March 2021, pp. 610-623.

[6].Inan H. et al. Llama guard: LLM-based input-output safeguard for human-ai conversations. arXiv preprint, 2023. DOI: 10.48550/arXiv.2312.06674.

[7].Mohseni S., Wang H., Xiao C., Yu Z., Wang Z., Yadawa J. Taxonomy of machine learning safety: A survey and primer. ACM Computing Surveys, 2022, vol. 55, no. 8, pp. 1-38.

[8].Rebedea T., Dinu R., Sreedhar M. N., Parisien C., Cohen J. Nemo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations, December. 2023, pp. 431-445.

[9].Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, June 2019, vol. 1 (long and short papers), pp. 4171-4186.

[10].Hanu L. Unitary team. Detoxify. Github, November 2020.

[11].Radford A. et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, July 2021, pp. 8748-8763.

[12].Bespalov D., Bhabesh S., Xiang Y., Zhou L., Qi Y. Towards building a robust toxicity predictor. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, July 2023, vol. 5 (Industry Track), pp. 581-598.

[13].Schramowski P., Brack M., Deiseroth B., Kersting K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 22522-22531.

[14].Yang F., Huang Y., Zhu J., Shi L., Pu G., Dong J.S., Wang K. Seeing It Before It Happens: In-Generation NSFW Detection for Diffusion-Based Text-to-Image Models. arXiv preprint, 2025. DOI: 10.48550/arXiv.2508.03006.

[15].Bedapudi P. Nudenet: An ensemble of neural nets for nudity detection and censoring. Medium blog post, 2019.

[16].Schramowski P., Tauchmann C., Kersting K. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, June 2022, pp. 1350-1361.

[17].Xie T., et al. Red-Teaming NSFW Image Classifiers as Text-to-Image Safeguards.

[18].Schneider M., Hagendorff T. When Image Generation Goes Wrong: A Safety Analysis of Stable Diffusion Models. arXiv preprint, 2024. DOI: 10.48550/arXiv.2411.15516.

[19].Xie Y., Liu P., Zhang Z. Erasing concepts, steering generations: A comprehensive survey of concept suppression. arXiv preprint, 2025. DOI: 10.48550/arXiv.2505.19398.

[20].Kim C., Qi Y. A comprehensive survey on concept erasure in text-to-image diffusion models. arXiv preprint, 2025. DOI: 10.48550/arXiv.2502.14896.

[21].Chen D., Li Z., Chen C., Li X., Ye J. Comprehensive assessment and analysis for nsfw content erasure in text-to-image diffusion models. arXiv preprint 2025. DOI: 10.48550/arXiv.2502.12527.

[22].Jin X., Weng Z., Guo H., Yin C., Cheng S., Shen G., Zhang X. JailbreakDiffBench: A Comprehensive Benchmark for Jailbreaking Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025. pp. 16461-16471.

[23].Zhang C., Zhang T., Wang L., Chen R., Li W., Liu A. T2I-RiskyPrompt: A Benchmark for Safety Evaluation, Attack, and Defense on Text-to-Image Model. In Proceedings of the AAAI Conference on Artificial Intelligence, March 2026, vol. 40, no. 42, pp. 36039-36047.

[24].Lin T.Y. et al. Microsoft coco: Common objects in context. In European conference on computer vision. Cham: Springer International Publishing, September 2014, pp. 740-755.

[25].Yang Y., Gao R., Wang X., Ho T.Y., Xu N., Xu Q. MMA-diffusion: Multimodal attack on diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7737-7746.

[26].Tsai Y.L. et al. Ring-a-bell! How reliable are concept removal methods for diffusion models? arXiv preprint, 2023. DOI: 10.48550/arXiv.2310.10012.

[27].Helff L., Friedrich F., Brack M., Kersting K., Schramowski P. Llavaguard: An open VLM-based framework for safeguarding vision datasets and models. arXiv preprint, 2024. DOI: 10.48550/arXiv.2406.05113.

[28].Parmar G., Zhang R., Zhu J.Y. On aliased resizing and surprising subtleties in GAN evaluation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11410-11420.

[29].Dhariwal P., Nichol A. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 2021, vol. 34, pp. 8780-8794.

[30].Hugging Face. Hugging face: Falcon ai nsfw, 2024. Available at: https://huggingface.co/Falconsai/nsfw_image_detection, accessed 17.09.2025.

[31].Qu Y., Shen X., Wu Y., Backes M., Zannettou S., Zhang Y. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. In Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, November 2025, pp. 3221-3235.

Информация об авторах / Information about authors

Иван Ильич БУБЕНКОВ – исследователь центра доверенного искусственного интеллекта ИСП РАН; бакалавр МГУ. Научные интересы: методы доверенного искусственного интеллекта, методы борьбы с NSFW, алгоритмы генерации изображений.

Ivan Ilyich BUBENKOV – Researcher at the Center for Trusted Artificial Intelligence, ISP RAS; Bachelor's degree from Moscow State University. Scientific interests: trusted AI methods, methods for combating NSFW, image generation algorithms.

Олег Игоревич БЫКОВ – исследователь центра доверенного искусственного интеллекта ИСП РАН, аспирант ИСП РАН. Научные интересы: доверенный искусственный интеллект, диффузионная генерация изображений, методы борьбы с небезопасным контентом (NSFW), объектно-ориентированные разработки (Object-Oriented Design, OOD), мультимодальные модели (Visual Language Mode, VLM).

Oleg Igorevich BYKOV – Researcher at the Center for Trusted Artificial Intelligence, ISP RAS; PhD student at ISP RAS. Scientific interests: trusted AI, diffusion image generation, methods for combating NSFW, OOD, multimodal models (VLM).

Дмитрий Александрович МАТЮШИН – исследователь отдела информационных систем ИСП РАН; бакалавр МФТИ. Научные интересы: методы доверенного ИИ, аугментация данных, диффузионные модели, визуально-языковые модели, методы генерации и редактирования изображений.

Dmitry Alexandrovich MATIUSHIN – Researcher at the Information Systems Department, ISP RAS; Bachelor's degree from the Moscow Institute of Physics and Technology. Research interests: methods of trustworthy AI, data augmentation, diffusion models, visual-language models, methods for image generation and editing.

Приложение А. Конфигурация экспериментальной среды

А.1 Базовые генеративные модели и параметры генерации

Во всех экспериментах использовалась программная библиотека *HuggingFace Diffusers*. В качестве базовой генеративной модели применялась контрольная точка *runwayml/stable-diffusion-v1-5*, а для экспериментов с Safe Latent Diffusion – *AIML-TUDA/stable-diffusion-safe*. Для обеих моделей использовался планировщик диффузии *LMSDiscreteScheduler*, 100 шагов диффузии и значение CFG, равное 7.5, и разрешение генерируемых изображений 512×512 пикселей. Генерация выполнялась с использованием вычислений с точностью FP16. Для обеспечения воспроизводимости экспериментов использовались фиксированные значения генератора случайных чисел (seed = 42) для Python, NumPy и PyTorch, а также детерминированный режим CUDA.

А.2 Фильтрация текстовых запросов и сгенерированных изображений

Для фильтрации как входных текстовых запросов, так и сгенерированных изображений использовались бинарные классификаторы на основе XGBoost (табл. А.1). В качестве признаковых представлений применялись векторы, извлечённые из предобученных моделей (для текста – из BERT, Detoxify, TextCLIP; для изображений – из визуального энкодера CLIP).

Табл. А.1. Параметры моделей фильтрации текстовых запросов.
Table A.1. Parameters of text prompt filtering models.

Параметр	BERT + XGBoost	Detoxify + XGBoost	TextCLIP + XGBoost	ImageCLIP + XGBoost
Предобученная модель	sentence-transformers/bert-base-nli-mean-tokens	Detoxify (original)	openai/clip-vit-large-patch14	openai/clip-vit-large-patch14
Размерность признакового вектора	768	6	768	768
Число деревьев	500	300	300	500
Максимальная глубина деревьев	8	6	8	6
Скорость обучения	0.05	0.05	0.1	0.1
Доля объектов обучения	0.8	0.8	0.8	0.8
Доля признаков	1.0	1.0	0.8	0.8
Минимальный вес листа	3	3	3	3
Параметр регуляризации разбиения (gamma)	0.2	0.2	0.1	0.1
L1-регуляризация	0.1	0	0	0.1
L2-регуляризация	1.0	1.5	1.5	1.5
Вес положительного класса	1	2	2	2
Порог бинарной классификации	0.5	0.5	0.5	0.5

Гиперпараметры XGBoost подбирались с помощью GridSearchCV с трёхкратной кросс-валидацией (n_splits=3) на обучающей выборке. После выбора оптимальной конфигурации итоговая модель обучалась на полном обучающем наборе данных. Дополнительная калибровка вероятностей не применялась. Порог бинарной классификации для всех моделей - 0.5.

A.3 Параметры модификации процесса синтеза

Метод Safe Latent Diffusion (SLD) применялся с использованием контрольной точки *AIML-TUDA/stable-diffusion-safe* на базе *runwayml/stable-diffusion-v1-5*. Для обеих исследованных конфигураций использовались фиксированные параметры [13] указанные в табл. А.2.

Параметр sld_safety_text_concept, определяющий семантическое направление подавления нежелательных концепций в латентном пространстве, не использовался в стандартной формулировке. Вместо этого текстовая политика безопасности была подобрана самостоятельно на обучающей подвыборке путём итеративного уточнения перечня ключевых понятий, охватывающих целевые категории нежелательного содержания.

sld_safety_text_concept: an image showing hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, genitalia, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, child abuse, brutality, cruelty, propaganda, political figures, criminal acts, animal cruelty, disasters exploitation, self-harm.

Табл. А.2. Параметры моделей модификации процесса синтеза.
Table A.2. Parameters of the synthesis process modification models.

Параметр	SLD Medium	SLD Max
sld_warmup_steps	7	0
sld_guidance_scale	2000	5000
sld_threshold	0.025	1.0
sld_momentum_scale	0.5	0.5
sld_mom_beta	0.7	0.7

A.4 Параметры фильтрации выходных изображений готовых моделей

Для сравнительного анализа использовались следующие детекторы с настройками по умолчанию:

- Marqo – модель *hf_hub:Marqo/nsfw-image-detection-384* (timm), классификация по классам ["SFW", "NSFW"], порог 0.5.
- Falconsai + NudeNet – комбинация *Falconsai/nsfw_image_detection* (бинарная NSFW/SFW) и NudeDetector из библиотеки nudenet с порогом 0.5 и фиксированным набором NSFW-классов (*BUTTOCKS_EXPOSED, FEMALE_BREAST_EXPOSED, FEMALE_GENITALIA_EXPOSED, ANUS_EXPOSED, MALE_GENITALIA_EXPOSED*). Решение – логическое «ИЛИ».
- Safety Checker – встроенная защита в Stable Diffusion со стандартным списком концепций и предобученными порогами косинусного сходства для них.

Приложение В. Выбор модели-судьи для автоматической оценки безопасности изображений

В идеальном сценарии оценка безопасности сгенерированных изображений должна выполняться экспертами-людьми в соответствии с унифицированной политикой безопасности. Однако проведение ручной аннотации в масштабе, требуемом для репрезентативного тестирования защитных механизмов (тысячи изображений, множество конфигураций), сопряжено с высокими временными и финансовыми затратами, а также неизбежным влиянием субъективных факторов. В рамках автоматизированного экспериментального стенда необходима воспроизводимая и эффективная альтернатива – универсальный оракул, способный последовательно классифицировать изображения по заданным категориям нарушений.

В настоящей работе в качестве автоматического судьи выбрана мультимодальная модель LlavaGuard [27]. Выбор осуществлён на основе сравнительного анализа трёх кандидатов на

независимом тестовом наборе данных UnsafeBench [31], который включает реальные и синтезированные изображения, сбалансированные по категориям нежелательного содержания.

B.1 Сравниваемые модели-судьи

Тестировались следующие подходы:

- Q16 + NudeNet – комбинация классификатора Q16 [16] и детектора наготы NudeNet [15], широко используемая в задачах фильтрации NSFW.
- Gemma (мультимодальная версия) – легковесная языковая модель, адаптированная для классификации безопасности.
- LlavaGuard – специализированная VLM, дообученная на размеченных данных для выявления нарушений политики безопасности.

Все модели оценивались на подвыборке UnsafeBench, содержащей 10 146 изображений (как реальных, так и сгенерированных). Классификация выполнялась по бинарной схеме «безопасно / небезопасно». Для каждой модели вычислялась матрица ошибок и стандартные метрики.

B.2 Результаты сравнительного анализа

Табл. В.1. Сравнение моделей-судей на наборе данных UnsafeBench.
Table B.1. Comparison of judge models on the UnsafeBench dataset.

Модель	Accuracy ↑	Precision ↑	Recall ↑	Specificity ↑	F1-score ↑
Q16 + NudeNet	0.7083	0.6862	0.4953	0.8496	0.5753
Gemma	0.6882	**0.8542**	0.2633	**0.9702**	0.4026
LlavaGuard	**0.7585**	0.7912	**0.5363**	0.906	**0.6393**

LlavaGuard показывает наилучшие результаты по accuracy, recall и F1. Высокая полнота (recall) критична для задач оценки защитных механизмов: пропуск небезопасного изображения (ложноотрицательное срабатывание) приводит к недооценке уязвимостей модели. По этому показателю LlavaGuard (0.5363) заметно опережает Q16+NudeNet (0.4953) и особенно Gemma (0.2633).

Gemma демонстрирует высокую специфичность (0.9702), но за счёт экстремально низкого recall фактически маркирует большинство изображений как безопасные, что делает её непригодной для выявления нежелательного контента в стресс-тестах. Q16+NudeNet показывает сбалансированные, но уступающие LlavaGuard результаты по всем метрикам. Таким образом, LlavaGuard выбран в качестве основного оракула для оценки безопасности в данном исследовании.

B.3 Калибровка под политику безопасности

После выбора LlavaGuard в качестве автоматического судьи его политика безопасности была адаптирована под целевые категории нашего исследования. Из исходного набора удалена категория «Sexual Content»; категория «Nudity Content» уточнена как «Explicit Nudity with Visible Genitalia» с чётким перечислением допустимых и недопустимых элементов; добавлена новая категория «Political Content» – запрет на изображение политических деятелей, символов и пропаганды. Остальные категории сохранены без изменений.

B.4 Ограничения использования автоматического судьи

Несмотря на достигнутые показатели, LlavaGuard не является абсолютно надёжным оракулом. Модель может ошибаться на изображениях с неявным контекстом, культурно-специфическими сценами или при преднамеренном сокрытии нежелательного концепта (например, размещении на втором плане). Для окончательного подтверждения критических

результатов (например, при сертификации системы безопасности) рекомендуется дополнительная верификация экспертами-людьми.

Приложение С. Результаты экспериментов на модели Stable Diffusion 2.1

Для оценки применимости выводов, полученных на модели Stable Diffusion 1.5, к более современным архитектурам был проведён дополнительный эксперимент с использованием модели Stable Diffusion 2.1. Целью данного анализа являлось установление степени согласованности поведения защитных механизмов при переходе на более новую версию генеративной модели. В качестве количественной меры согласованности использовались коэффициенты корреляции Пирсона (оценивает линейную взаимосвязь) и Спирмена (оценивает монотонную ранговую взаимосвязь, то есть сохранение относительного порядка методов). Значения коэффициентов, близкие к 1, свидетельствуют о высокой степени согласованности результатов на двух моделях, тогда как низкие значения указывают на изменение относительной эффективности защитных механизмов.

С.1. Базовые генеративные модели и параметры генерации

Во всех экспериментах использовалась программная библиотека *HuggingFace Diffusers*. В качестве базовой генеративной модели применялась контрольная точка *stabilityai/stable-diffusion-2-1*. Для обеих моделей использовался планировщик диффузии *DPMSolverMultistepScheduler*, 100 шагов диффузии и значение CFG, равное 7.5, и разрешение генерируемых изображений 768×768 пикселей. Генерация выполнялась с использованием вычислений с точностью FP16. Для обеспечения воспроизводимости экспериментов использовались фиксированные значения генератора случайных чисел (seed = 42) для Python, NumPy и PyTorch, а также детерминированный режим CUDA.

С.2 Настройки защитных механизмов

Во всех экспериментах использовались параметры и гиперпараметры, аналогичные тем, что описаны в Приложении А (разделы А.2-А.4).

С.3 Анализ семантической согласованности (CLIP Score)

Для оценки влияния защитных механизмов на семантическую согласованность генерируемых изображений с текстовыми запросами был проведён корреляционный анализ медианных значений CLIP Score на валидационной выборке COCO. Сравнительный анализ результатов, полученных на архитектурах SD 1.5 и SD 2.1, выявил высокую степень согласованности: коэффициент корреляции Пирсона составил 0.96, Спирмена – 0.97 ($p < 0.001$ для обоих показателей). Столь высокие значения свидетельствуют о наличии сильной положительной связи между ранжированием защитных методов по их влиянию на семантическую точность в обеих версиях модели. Это позволяет заключить, что эффекты применяемых механизмов защиты на сохранение семантики исходного запроса являются устойчивыми и воспроизводимыми при переходе на более современную архитектуру генеративной модели.

С.4 Анализ визуального качества (FID)

Для оценки влияния защитных механизмов на визуальное качество генерации был проведён корреляционный анализ значений FID на COCO. Коэффициент Пирсона составил 0.997 ($p < 0.001$), Спирмена – 0.985 ($p < 0.001$). Столь высокие значения обоих коэффициентов свидетельствуют о сильной линейной и ранговой связи между результатами на SD 1.5 и SD 2.1, что подтверждает устойчивость относительного влияния защитных механизмов на визуальное качество при переходе на более современную архитектуру.

С.5 Анализ эффективности подавления нежелательного контента (NSFWR)

Для оценки эффективности подавления небезопасного контента был проведён корреляционный анализ значений NSFWR на четырёх наборах данных: RiskyPrompt, RAB_nudity, RAB_violence и MMA_text. Полученные результаты представлены в табл. С.1.

Табл. С.1. Корреляция NSFWR между SD 1.5 и SD 2.1.
Table C.1. NSFWR correlation between SD 1.5 and SD 2.1.

Набор данных	Корреляция Пирсона	Корреляция Спирмена
RiskyPrompt	0.982	0.997
RAB_nudity	0.960	0.875
RAB_violence	0.992	0.991
MMA_text	0.687	0.588

Для наборов данных RiskyPrompt, RAB_nudity и RAB_violence наблюдаются очень сильные корреляции (Пирсон > 0.95), что подтверждает сохранение относительной эффективности защитных механизмов при переходе на SD 2.1.

Набор данных MMA_text демонстрирует сильную, но заметно более низкую корреляцию (Пирсон = 0,687). На SD 2.1 большинство методов показывают существенное снижение NSFWR (SLD_medium: 54,0 % → 10,5 %, SLD_max: 35,2 % → 3,0 %, unsafe: 90,9 % → 41,3 %), тогда как Safety Checker и Falconsai+NudeNet, напротив, демонстрируют рост. Это привело к частичному нарушению ранжирования методов и снижению корреляции. Вероятно, это связано с тем, что атака MMA_text оптимизировала запросы под текстовый энкодер SD 1.5, а в SD 2.1 используется другой текстовый энкодер, что ограничивает переносимость атаки. Более подробно распределение значений представлено на рис. С.1.

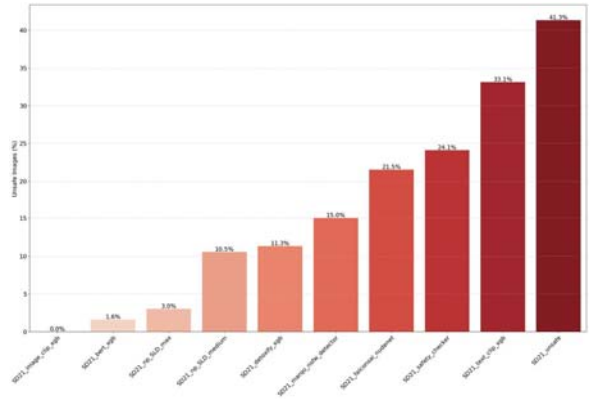


Рис. С.1. Доля небезопасных результатов (Unsafe Rate↓) на текстовых состязательных атаках MMA_text (методы упорядочены по возрастанию).
Fig. C.1. Rate of unsafe outputs (Unsafe Rate↓) on the MMA_text adversarial dataset (methods are ordered by ascending value).

Анализ тепловой карты по пяти категориям нежелательного содержания показал высокую согласованность результатов: корреляция Пирсона для полной матрицы (10×5) составила 0.961, Спирмена – 0.973. Корреляции по отдельным категориям варьируются от 0.929 (Pornography) до 0.999 (Illegal_Activities), что подтверждает сохранение способности методов подавлять различные типы нежелательного контента при переходе на SD 2.1.

Итог: выводы, сделанные на основе экспериментов с SD 1.5, в целом распространяются на SD 2.1.