DOI: 10.15514/ISPRAS-2026-38(5)-10

CC BY

Обратимые малошаговые диффузионные модели для редактирования изображений по текстовому описанию

Н.О. Стародубцев, ORCID: 0000-0002-5170-5989 <nstarodubtsev@hse.ru>

НИУ Высшая школа экономики,
101978, Россия, г. Москва, ул. Мясницкая, д. 20.

Аннотация. Дистилляция диффузионных моделей является одним из наиболее многообещающих подходов к синтезу высококачественных изображений, основанных на текстовых описаниях, с минимальным количеством шагов генерации. Однако, несмотря на успехи последних лет, существующие дистиллированные модели по-прежнему не обладают всем спектром возможностей диффузионных моделей – в частности, инверсией реальных изображений, которая лежит в основе многих точных методов редактирования изображений. Настоящая работа посвящена разработке эффективных методов кодирования реальных изображений в скрытое пространство с использованием дистиллированных диффузионных моделей. Для достижения этой цели мы предлагаем обратимую дистилляцию согласованности (invertible Consistency Distillation, iCD) – обобщённый подход дистилляции согласованности, обеспечивающий как высококачественный синтез изображений, так и качественное кодирование изображений всего за 3–4 прогона нейронной сети. Хотя задача инверсии для диффузионных моделей усложняется при высоких значениях управления без классификатора (classifier-free guidance), мы показываем, что динамическое управление (dynamic guidance) существенно снижает ошибку реконструкции без заметной потери качества генерации. В результате мы демонстрируем, что iCD в сочетании с динамическим управлением может служить высокоэффективным инструментом редактирования изображений по текстовому описанию без дополнительного дообучения (zero-shot), конкурируя с более затратными современными альтернативами.

Ключевые слова: текстово-управляемое редактирование изображений; дистилляция согласованности; диффузионные модели; инверсия изображений.

Для цитирования: Стародубцев Н.О. Обратимые малошаговые диффузионные модели для редактирования изображений по текстовому описанию. Труды ИСП РАН, 2026, том 38, вып. 5, стр. 155–174. DOI: 10.15514/ISPRAS-2026-38(5)-10.

Invertible few-step diffusion models for text-guided image editing

N.O. Starodubcev, ORCID: 0000-0002-5170-5989 <nstarodubtsev@hse.ru>

National Research University, Higher School of Economics,
20, Myasnitskaya st., Moscow, 101978, Russia,

Abstract. Diffusion distillation represents a highly promising direction for achieving faithful text-to-image generation in a few sampling steps. However, despite recent successes, existing distilled models still do not provide the full spectrum of diffusion abilities, such as real image inversion, which enables many precise image manipulation methods. This work aims to enrich distilled text-to-image diffusion models with the ability to effectively encode real images into their latent space. To this end, we introduce invertible Consistency Distillation (iCD), a generalized consistency distillation framework that facilitates both high-quality image synthesis and accurate image encoding in only 3–4 inference steps. Though the inversion problem for text-to-image diffusion models gets exacerbated by high classifier-free guidance scales, we notice that dynamic guidance significantly reduces reconstruction errors without noticeable degradation in generation performance. As a result, we demonstrate that iCD equipped with dynamic guidance may serve as a highly effective tool for zero-shot text-guided image editing, competing with more expensive state-of-the-art alternatives.

Keywords: text-guided image editing; consistency distillation; diffusion models; image inversion.

For citation: Starodubcev N.O. Invertible few-step diffusion models for text-guided image editing. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 5, pp. 155-174 (in Russian). DOI: 10.15514/ISPRAS-2026-38(5)-10.

1. Введение

В последние годы диффузионные модели для генерации изображений по текстовому описанию [1-6] стали доминирующей парадигмой синтеза изображений на основе текстовых запросов пользователя. Исключительное качество таких моделей делает их ценным инструментом для графических редакторов, в частности для решения различных задач редактирования изображений [7-9]. Однако на практике применимость диффузионных моделей часто ограничена их низкой скоростью вывода, обусловленной последовательной процедурой генерации, постепенно восстанавливающей изображение из случайного шума.

Чтобы ускорить процесс вывода, в последнее время проводится множество исследований, направленных на сокращение количества диффузионных шагов с помощью пошаговой дистилляции [10-17]. Эта методика показала впечатляющие результаты в области генерации высококачественных изображений за 1-4 шага и была успешно применена к современным моделям генерации изображений по тексту [18-24]. Хотя современные методы дистилляции зачастую предполагают компромисс между полнотой охвата мод распределения и качеством изображения в пользу небольшого количества шагов для получения результата, предложенные модели уже продемонстрировали свою эффективность в практических приложениях, таких как редактирование изображений по текстовым описаниям [25-27]. Наиболее эффективные методы редактирования на основе диффузионных моделей, как правило, требуют кодирования реальных изображений в скрытое пространство диффузионной модели. Для «недистиллированных» моделей такое кодирование возможно благодаря связи стохастического дифференциального уравнения (СДУ) и обыкновенного дифференциального уравнения вероятностного потока (probability flow ODE, PF ODE) [30]. Взгляд на диффузионные модели через призму ОДУ раскрывает их *обратимость (reversibility)*, то есть способность кодировать реальное изображение в скрытое пространство модели и затем восстанавливать его с минимальными изменениями. Эта способность успешно используется в различных приложениях (рис. 1), таких как редактирование изображений по текстовому описанию [31-33], перенос между доменами (domain translation) [34, 9], перенос стиля [35].

Рис. 1. Обратимая дистилляция согласованности (iCD) обеспечивает как быстрое редактирование изображений (порядка 0.9 секунд на изображение), так и высокое качество генерации при малом числе обращений к модели.

Fig. 1. Invertible Consistency Distillation (iCD) enables both fast image editing (about 0.9 seconds per image) and strong generation performance in a few model evaluations.

Однако остаётся вопрос: можно ли сделать дистиллированные модели обратимыми? Существующие методы дистилляции диффузии в основном сосредоточены на том, чтобы эффективно генерировать изображения. В этой работе мы отвечаем на этот вопрос утвердительно, предлагая обратимую дистилляцию согласованности (invertible Consistency Distillation, iCD). iCD – это обобщённый подход, который обеспечивает как высококачественную генерацию изображений, так и точную инверсию с небольшим количеством шагов обращения к модели.

На практике модели генерации изображений по тексту используют управление без классификатора (classifier-free guidance, CFG) [36], которое критически важно для соответствия между текстом и изображением [1, 3] и для редактирования по текстовому описанию [32-33]. Однако применение управляемых диффузионных процессов вызывает серьёзные трудности для методов редактирования, основанных на инверсии [32]. Предыдущие подходы [25-27, 32, 37-44] подробно рассматривали эти проблемы, но зачастую требуют значительных вычислительных ресурсов для достижения как масштабных изменений в изображении, так и точного сохранения его содержания. Хотя некоторые из этих методов можно применять и к дистиллированным моделям [25-27], они не позволяют в полной мере реализовать основное преимущество дистиллированных диффузионных моделей – эффективность вывода.

Один из ключевых элементов подхода iCD – то, как он работает с направляемыми диффузионными процессами. Недавно было предложено динамическое управление (dynamic guidance) для улучшения разнообразия распределения без заметной потери качества изображения [45-46]. Основная идея заключается в отключении CFG при высоких уровнях диффузионного шума, чтобы стимулировать исследование пространства решений на более ранних шагах генерации. В этой работе мы показываем, что динамический CFG способен улучшать инверсию изображений, сохраняя при этом возможность редактирования диффузионных моделей генерации изображений по тексту. Примечательно, что динамический CFG не создаёт дополнительных вычислительных затрат, полностью сохраняя выигрыш в эффективности от дистилляции диффузии. В наших экспериментах мы

показываем, что обратимые дистиллированные модели в сочетании с динамическим управлением являются высокоэффективным инструментом редактирования изображений на основе инверсии.

2. Вклад

- Мы предлагаем обобщённый подход дистилляции согласованности – обратимую дистилляцию согласованности (iCD), обеспечивающую как быструю генерацию изображений по тексту, так и быстрое кодирование изображений примерно за 3–4 шага обращения к нейронной сети.
- Мы исследуем динамическое управление без классификатора в контексте инверсии изображений и редактирования по текстовому описанию. Мы показываем, что оно сохраняет возможность редактирования диффузионных моделей генерации изображений по тексту, при этом существенно повышая качество инверсии без дополнительных затрат.
- Мы применяем iCD к крупномасштабным моделям генерации изображений по тексту, таким как Stable Diffusion 1.5 [4] и SDXL [1], и проводим их всестороннюю оценку для задач редактирования изображений. Результаты наших исследований, как автоматизированных, так и пользовательских, свидетельствуют о том, что iCD позволяет редактировать изображения по текстовым описаниям всего за 6–8 шагов. Это сравнимо с современными методами редактирования изображений по тексту, но при этом в 2–20 раз быстрее.

3. Обзор литературы

Диффузионные вероятностные модели. Диффузионные вероятностные модели (DPM) [28, 29, 47] – это класс порождающих моделей, генерирующих объекты из простого, как правило стандартного нормального, распределения путём решения соответствующего обыкновенного дифференциального уравнения вероятностного потока (Probability Flow ODE) [30, 48], что предполагает итеративное оценивание *score-функции*. DPM обучаются аппроксимировать score-функцию и используют специализированные решатели диффузионного ОДУ [48–50] для генерации. DDIM [49] – простой, но эффективный решатель, широко применяемый в моделях генерации изображений по тексту и работающий примерно за 50 шагов. Один шаг DDIM от x_t к x_s можно записать как

$$x_s^w = DDIM(x_t, t, s, c, w) = \sqrt{\frac{\alpha_s}{\alpha_t}} x_t + \epsilon_\theta^w(x_t, t, c)\left(\sqrt{1-\alpha_s} - \sqrt{\frac{\alpha_s}{\alpha_t}}\sqrt{1-\alpha_t}\right), \qquad (1)$$

где α_s, α_t определяются диффузионным расписанием [47], а $\epsilon_\theta^w(x_t, t, c) = \epsilon_\theta(x_t, t, \varnothing) + w\big(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \varnothing)\big)$ – линейная комбинация условного и безусловного предсказаний шума, называемая управлением без классификатора (Classifier-Free Guidance, CFG) [51] и используемая для повышения качества изображения и его соответствия текстовому условию при условной генерации. Далее для простоты мы опускаем условие c.

Благодаря обратимости PF ODE может решаться в обоих направлениях: кодирование данных в пространство шума и декодирование обратно без каких-либо дополнительных процедур оптимизации. Мы называем этот процесс *инверсией* [32], где кодирование и декодирование соответствуют *прямому* (forward) и *обратному* (reverse) процессам.

Дистилляция согласованности. Дистилляция согласованности (Consistency Distillation, CD) [10, 12, 13] – современный подход к дистилляции диффузии для генерации изображений за малое число шагов, обучающий модель интегрировать PF ODE, порождённое предобученной диффузионной моделью. Конкретней, модель f_θ обучается удовлетворять свойству самосогласованности:

$$L_{CD}(\theta) = E\left[d\left(f_\theta\left(x^w_{t_{n-1}}, t_{n-1}\right), f_\theta\left(x_{t_n}, t_n\right)\right)\right] \to \min_\theta, \quad (2)$$

где $t_n \in \{t_0, \ldots, t_N\}$ – дискретный временной шаг, $d(\cdot,\cdot)$ – функция расстояния, а $x^w_{t_{n-1}}$ получается одним шагом решателя DDIM от t_n к t_{n-1} с использованием диффузионной модели-учителя. Оптимум (2) определяется граничным условием $f_\theta\left(x_{t_0}, t_0\right) = x_{t_0}$. Таким образом, модели согласованности (consistency models, CM) обучаются переходу из любой точки траектории в начальную точку: $f_\theta\left(x_{t_n}, t_n\right) = x_{t_0}$, $\forall\, t_n \in \{t_0, \ldots, t_N\}$. Следовательно, CM допускают генерацию за один шаг.

Динамическое управление. Современные модели генерации изображений по тексту используют большие значения CFG для достижения высокого качества изображения и точного соответствия текстовому запросу. Однако это часто приводит к снижению разнообразия генерируемых изображений. Чтобы решить эту проблему, недавно было предложено *динамическое управление без классификатора* (dynamic classifier-free guidance) [45, 46, 54], повышающее разнообразие распределения без заметной потери качества генерации. CADS [45] постепенно увеличивает значение управления от нуля до исходного высокого значения на протяжении процесса генерации (рис. 2a). Альтернативно, в работе [46] предлагается отключать управление при низких и высоких уровнях шума, используя его только на среднем интервале временных шагов (рис. 2b). Обе стратегии указывают на то, что именно ненаправляемый процесс при высоких уровнях шума отвечает за улучшение разнообразия распределения без потери качества генерации. Кроме того, авторы [46] показывают, что управление при низких уровнях шума оказывает незначительное влияние на качество и может быть опущено, чтобы избежать дополнительных обращений к модели для его расчёта. Обе динамические техники управляются двумя гиперпараметрами, τ_1 и τ_2, определяющими значение динамического CFG $w(t)$. В данной работе мы сосредоточимся на формулировке CADS.

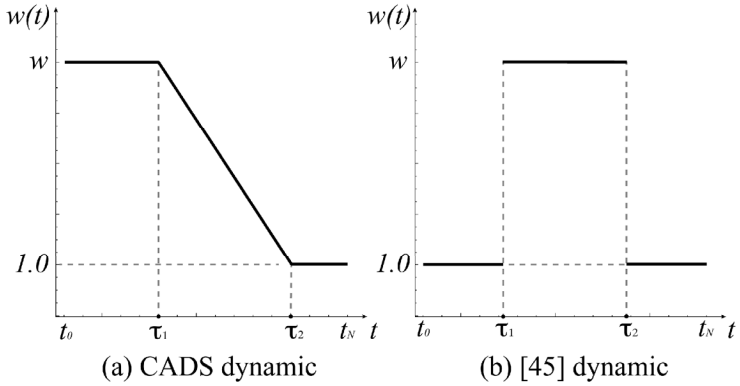

Рис. 2. Стратегии динамического CFG.
Fig. 2. Dynamic CFG strategies.

4. Метод

В этом разделе представлен подход обратимой дистилляции согласованности (iCD), включающий прямую и обратную модели согласованности. Сначала мы формулируем процедуру прямой CD, кодирующую изображения в скрытое представление шума. Затем мы описываем многограничное обобщение iCD, раскрывающее детерминированную многошаговую инверсию. Наконец, мы исследуем технику *динамического управления* с точки зрения задачи инверсии.

4.1 Прямая дистилляция согласованности

Прямая дистилляция согласованности (forward Consistency Distillation, fCD) работает противоположно CD: она нацелена на отображение произвольной точки траектории PF ODE в скрытое представление шума (последнюю точку траектории).

Переход от CD к прямому аналогу довольно прост: единственное, что требуется изменить, – это граничное условие. А именно, прямая модель согласованности ограничивается требованием быть тождественной функцией для последней точки траектории: $f_\theta\left(x_{t_N}, t_N\right) = x_{t_N}$. Таким образом, fCD наследует ту же функцию потерь дистилляции согласованности (2), не требуя дополнительных затрат на обучение. Благодаря этому дистиллированная модель может преобразовывать любую точку траектории в последнюю: $f_\theta\left(x_{t_n}, t_n\right) = x_{t_N}$, $\forall\, t_n$. Чтобы выполнить инверсию, сначала fCD кодирует изображение в шум, а затем CD декодирует его обратно. Сравнение CD и fCD показано на рис. 3a.

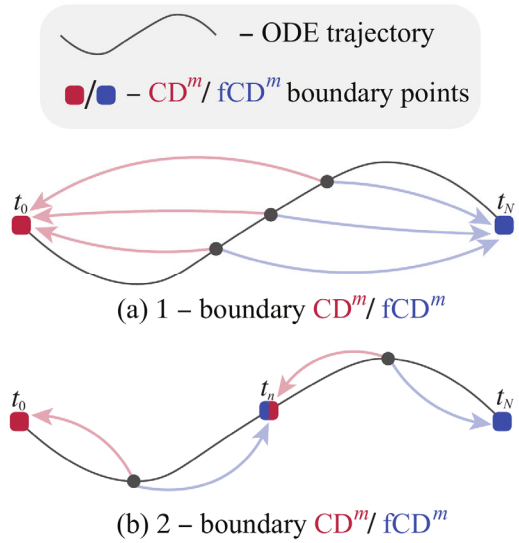

Рис. 3. Предложенный подход обратимой дистилляции согласованности состоит из двух моделей: прямой m-граничной модели fCD^m и обратной модели CD^m. (a) При $m = 1$ обратная модель соответствует CD. Большее число граничных точек раскрывает детерминированную многошаговую инверсию, например (b) показывает случай $m = 2$.
Fig. 3. The proposed invertible Consistency Distillation framework consists of two models: the forward m-boundary model, fCD^m, and the reverse model, CD^m. (a) For $m = 1$, the reverse model corresponds to CD. More boundary points unlock the deterministic multistep inversion, e.g., (b) shows the case for $m = 2$.

4.2 Многограничная дистилляция согласованности

На практике кодирование с помощью fCD имеет две основные проблемы. Во-первых, как и в случае CD, предсказание fCD за один шаг может быть весьма неточным. Однако эту проблему нельзя легко решить, поскольку *стохастическая многошаговая процедура генерации* [10] неприменима к fCD: промежуточные точки траектории не могут быть получены из скрытого представления шума с помощью прямого диффузионного процесса. Во-вторых, даже если fCD точна, *стохастическая многошаговая процедура генерации* не подходит для декодирования, поскольку её стохастическая природа не позволяет

восстанавливать реальные изображения. Поэтому для повышения точности предсказаний fCD и снижения ошибки реконструкции CD необходимо сформулировать детерминированную многошаговую процедуру для обеих моделей.

Недавние подходы [13, 53] обобщают подход CD на многошаговый режим, позволяя аппроксимировать произвольные интервалы траектории в обратном направлении. Однако эти методы сфокусированы исключительно на генерации и не поддерживают инверсию. Опираясь на эти работы, мы предлагаем многограничную CD, которая раскрывает детерминированную многошаговую инверсию у дистиллированных моделей и требует затрат на обучение, сопоставимых с классическими методами CD.

А именно, мы разбиваем интервал генерации $\{t_0, \dots, t_N\}$ на m сегментов и выполняем дистилляцию отдельно на каждом из них. Таким образом мы получаем набор одношаговых моделей согласованности, работающих на разных интервалах и с разными граничными точками. Такая формулировка справедлива как для CD, так и для fCD, и, следовательно, позволяет реализовать детерминированную многошаговую инверсию. Иллюстрация 2-граничной CD и fCD представлена на рис. 3b. Мы обозначаем многограничные обратную и прямую модели как CD^m и fCD^m.

Формально мы рассматриваем CD^m и fCD^m, используя следующую параметризацию, вдохновлённую работами [13, 53]:

$$x_{s_t^m} = f_\theta^m(x_t, t, s_t^m, w) = DDIM(x_t, t, s_t^m, w), \qquad (3)$$

где s_t^m – граничный временной шаг, зависящий от числа границ m и текущего шага t. Например, пусть $m = 1$, тогда $s_t^1 = t_0$ для CD^1 и $s_t^1 = t_N$ для fCD^1. Отметим, что мы обучаем одну модель, а многошаговая генерация и кодирование достигается варьированием s_t^m на этапе вывода. Целевая функция обучения остаётся такой же, как (2), не требуя дополнительных затрат на обучение по сравнению с CD. Единственное ограничение состоит в том, что число сегментов и соответствующие граничные временные шаги должны быть заданы до начала обучения.

4.3 Обучение fCD^m и CD^m

Мы обучаем fCD^m и CD^m раздельно, инициализируя обе модели одной и той же моделью-учителем. Мы используем одну и ту же функцию потерь, но с единственной разницей в виде граничных временных шагов. Однако заметное отличие состоит в масштабе CFG, w. Для CD^m мы предварительно встраиваем управление в модель, следуя [20], чтобы использовать различные значения w на этапе генерации и избежать дополнительных обращений к модели. Для fCD^m мы рассматриваем модель без управления с постоянным $w = 1$. Причина в том, что направляемое кодирование ($w > 1$) приводит к скрытому представлению шума, выходящему за пределы обучающего распределения [32], и, как следствие, к плохой реконструкции изображения. Наконец, мы обнаруживаем, что $m = 3 - 4$ обеспечивает конкурентоспособное качество генерации и инверсии для крупномасштабных моделей генерации изображений по тексту [1, 4].

Функции потерь сохранения. Описанная выше процедура уже обеспечивает достаточно хорошее качество инверсии, но всё ещё не достигает качества инверсии модели-учителя. Чтобы сократить этот разрыв, мы предлагаем прямую и обратную функции потерь сохранения (preservation losses), нацеленные на повышение согласованности CD^m и fCD^m друг с другом и на улучшение точности инверсии. Эти функции потерь могут дополнительно включаться на этапе обучения. Далее мы обозначаем параметры CD^m и fCD^m как θ^+ и θ^- соответственно.

Прямая функция потерь сохранения изменяет только fCD^m и записывается следующим образом:

$$L_f(\theta^-, \theta^+) = E\left[d\left(f_{\theta^-}^m\left(f_{\theta^+}^m(x_{s_t^m})\right), x_{s_t^m}\right)\right] \to \min_{\theta^-}, \qquad (4)$$

Для простоты мы опускаем часть обозначений. Вкратце, мы случайным образом выбираем зашумлённое изображение $x_{s_t^m}$ для граничного временного шага s_t^m, затем делаем предсказание с помощью CD^m и побуждаем fCD^m предсказывать то же самое $x_{s_t^m}$. Такой подход улучшает согласованность между CD^m и fCD^m.

Обратная функция потерь сохранения основана на той же идее, но с той разницей, что оптимизируется другая модель (CD^m вместо fCD^m) и меняется порядок предсказаний: сначала делается предсказание с помощью fCD^m, а затем используется CD^m. Мы обозначаем её как $L_r(\theta^-, \theta^+)$. В наших экспериментах мы вычисляем функции потерь сохранения только для ненаправляемого обратного процесса ($w = 1$).

Итог. Мы представляем итоговый подход для случая, когда fCD^m и CD^m обучаются совместно, начиная с предобученной диффузионной модели. Однако возможно обучать их и по одной, беря уже предобученную модель согласованности и обучая только оставшуюся.

Итоговая целевая функция состоит из двух функций потерь согласованности с предложенной многограничной модификацией и двух функций потерь сохранения:

$$L_{iCD}(\theta^+, \theta^-) = L_{CD}(\theta^+) + L_{CD}(\theta^-) + \lambda_f L_f(\theta^-, \theta^+) + \lambda_r L_r(\theta^+, \theta^-). \qquad (5)$$

Таким образом, предложенный подход способен конкурировать с современными методами инверсии, использующими существенно более тяжеловесные диффузионные модели.

4.4 Динамическое управление для инверсии

Как обсуждалось выше, динамическое управление [45, 46] демонстрирует многообещающие результаты как для точной по содержанию, так и для разнообразной генерации изображений по тексту. В этой работе мы показываем, что динамический CFG также является эффективной техникой повышения точности инверсии, как показано на рис. 4b. Далее мы анализируем, как динамическое управление может способствовать инверсии изображений, сохраняя при этом качество генерации. Чтобы ответить на эти вопросы, мы проводим эксперименты с моделью Stable Diffusion 1.5 с решателем DDIM за 50 шагов и максимальным масштабом CFG, равным 8.0 [4]. Отметим, что модель Stable Diffusion 1.5 является «латентной» моделью, то есть выполняет диффузионный процесс не в пространстве пикселей, а в пространстве компактных скрытых, или латентных, представлений изображений, получаемых с помощью VAE [65].

Динамическое управление для декодирования. Мы начинаем с анализа динамического CFG на этапе декодирования, используя ненаправляемый процесс кодирования, следуя предыдущей работе [32]. Сначала нас интересует, на каких временных шагах направление оказывает наибольшее влияние на качество реконструкции. Для этого мы оцениваем среднеквадратичное отклонение (Mean Squared Error, MSE) между реальными и восстановленными изображениями для разных пороговых значений включения CFG T. Если $t > T$, мы устанавливаем $w = 1.0$, в противном случае масштаб CFG принимает исходное значение 8.0. На рис. 4a мы наблюдаем экспоненциальное снижение ошибки реконструкции, что означает, что отсутствие CFG при более высоких уровнях шума необходимо для более точной инверсии. Рис. 4b качественно подтверждает эту интуицию. Эти результаты согласуются с работами [45, 46], которые также предлагают отключать управление при высоких уровнях шума, но объясняют это с точки зрения улучшения разнообразия.

Затем мы исследуем влияние различных значений τ_1, τ_2 динамики CADS (рис. 2a) на качество инверсии и генерации. Мы стремимся найти рабочую точку, обеспечивающую как высокое качество генерации, так и точную по содержанию инверсию изображения. Для этого мы оцениваем качество генерации с помощью метрики ImageReward [55] (IR) на случайно сгенерированных образцах для 1000 текстовых запросов из COCO2014 [56]. Точность инверсии оценивается по MSE между исходными и восстановленными образцами. На рис. 5a представлены результаты для различных значений τ_1 и τ_2. Видно, что ряд точек при $\tau_1 \geq 0.7$ показывают немного более низкое качество генерации изображений по тексту, но заметно

лучшее качество реконструкции по сравнению с постоянным масштабом CFG, равным 8.0. Более того, мы замечаем, что настройки с $\tau_1 = \tau_2$ демонстрируют результаты, схожие с настройками при $\tau_1 < \tau_2$. Следовательно, во всех наших экспериментах мы рассматриваем единый параметр τ, представляющий случай $\tau_1 = \tau_2$, и используем значения $\tau = 0.7$ и $\tau = 0.8$. Это означает, что CD^m следует ненаправляемому декодированию при $t > \tau$ и устанавливает исходный масштаб CFG при $t \leq \tau$.

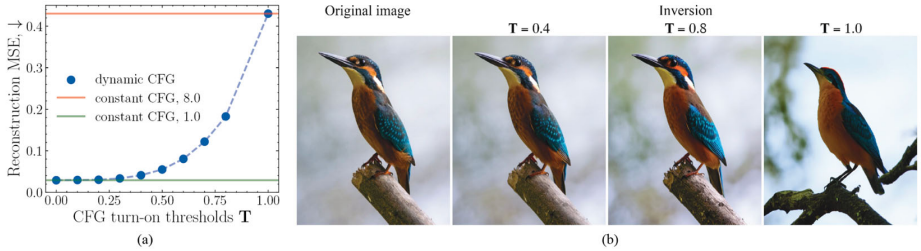

Рис. 4. (a) Ошибка реконструкции процесса декодирования для разных пороговых значений включения CFG. (b) Примеры инверсии изображений для разных пороговых значений включения CFG T. Направление при высоких уровнях шума ($T = 1.0$) резко ухудшает качество инверсии.

Fig. 4. (a) Reconstruction error of the decoding process for different CFG turn-on thresholds. (b) Image inversion examples for different CFG turn-on thresholds T. Guidance at high noise levels ($T = 1.0$) drastically degrades the inversion quality.

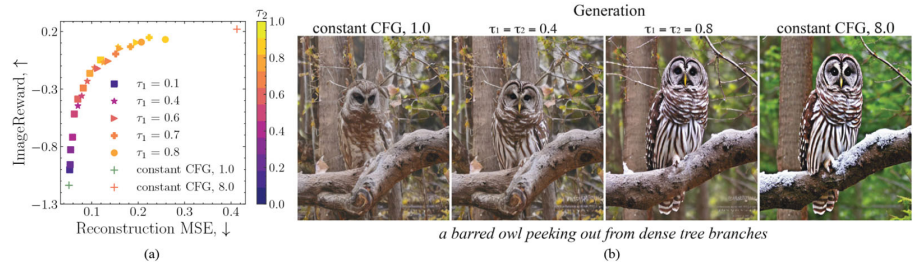

Рис. 5. (a) Компромисс между качеством генерации (IR) и качеством реконструкции (MSE), обеспечиваемый различными τ_1, τ_2. (b) Примеры генерации для динамического и постоянного масштабов CFG. Точки в районе $\tau_1 = \tau_2 = 0.8$ обеспечивают предпочтительный компромисс между качеством генерации и инверсии.

Fig. 5. (a) Trade-off between generation performance (IR) and reconstruction quality (MSE) provided by different τ_1, τ_2. (b) Generation examples for dynamic and constant CFG scales. The points around $\tau_1 = \tau_2 = 0.8$ provide preferable trade-off between generation and inversion performance.

Отметим, что настройка с $\tau = \tau_1 = \tau_2$ соответствует ступенчатой функции CFG $w(t)$, что даёт явное преимущество для дистиллированных моделей. Линейно изменяющиеся масштабы CFG неприменимы к процессам с крупным шагом дискретизации, типичным для дистиллированных диффузионных моделей. Поэтому подобное расписание CFG необходимо было бы дистиллировать в модель на этапе обучения, что снижает гибкость для разных задач генерации и редактирования. Напротив, ступенчатая функция CFG позволяет реализовать динамический CFG для уже предобученных дистиллированных моделей, работающих с разными постоянными масштабами CFG.

Динамическое управление для кодирования. Далее мы исследуем роль управления в решении задачи кодирования. В табл. 1 мы сравниваем скрытое представление шума, закодированные с использованием различных стратегий управления. Качество скрытых представлений шума оценивается по качеству генерации, начатой с этих представлений при

фиксированном масштабе CFG, равном 8. В качестве меры качества мы вычисляем FID для 5000 пар изображение-текст из COCO [56].

Мы наблюдаем, что скрытые представления, полученные при неизменно высоком масштабе CFG, показывают наихудшее качество генерации, что указывает на их выход за пределы области распределения. Хотя динамическое управление даёт заметно более правдоподобные представления, в большинстве случаев оно всё же уступает ненаправляемому кодированию. Чтобы дополнительно подтвердить эти результаты, мы оцениваем отрицательное логарифмическое правдоподобие (NLL) закодированных скрытых представлений при разных настройках CFG (табл. 1). NLL вычисляется относительно стандартного нормального распределения. В то время как NLL снижается для динамического CFG при уменьшении τ, кодирование без управления ($w = 1$) даёт наивысшее значение правдоподобия. Поэтому во всех наших экспериментах мы сохраняем $w = 1$ для кодирования и обучаем прямые дистиллированные модели (fCD) на ненаправляемом процессе учителя.

Табл. 1. FID-5k для SD1.5, начиная со скрытых представлений шума, полученных с помощью различных стратегий кодирования, и NLL (Negative Log-Likelihood) для этих представлений. Хотя кодирование с динамическим CFG стабильно даёт более правдоподобные представления по сравнению с постоянным CFG, ненаправляемое кодирование по-прежнему остаётся предпочтительным.

Table 1. FID-5k for SD1.5 starting from the noise latents obtained using different encoding strategies, and NLL for these latents. Though encoding with dynamic CFG produces consistently more plausible latents than constant CFG, the unguided encoding remains preferable.

	Кодирование без CFG	Кодирование с динамическим CFG, $\tau = 0.6$	Кодирование с динамическим CFG, $\tau = 0.8$	Кодирование с CFG
Декодирование без CFG	11.0	36.6	63.4	100.5
Декодирование с динамическим CFG, $\tau = 0.6$	11.4	**23.5**	67.8	102.6
Декодирование с динамическим CFG, $\tau = 0.8$	15.0	14.6	**52.2**	108.5
Декодирование с CFG	19.0	19.2	19.0	**102.1**
Latent NLL, ↓	1.401	1.409	1.415	1.428

5. Эксперименты

В приведённых далее экспериментах мы применяем наш подход к моделям генерации изображений по тексту разного масштаба: SD1.5 [57] и SDXL [1], обозначая их как iCD-SD1.5 и iCD-XL соответственно.

Сначала мы демонстрируем возможности инверсии предложенного подхода. Затем мы рассматриваем задачу редактирования изображений по текстовому описанию и показываем, что наш подход превосходит либо не уступает существенно более затратным базовым методам.

Прежде чем перейти к основным экспериментам, на рис. 6 мы приводим несколько сгенерированных образцов с использованием iCD-XL за 4 шага. Результаты подтверждают, что дистиллированная модель демонстрирует высокое качество генерации изображений по текстовому описанию.

5.1 Качество инверсии iCD

Здесь мы анализируем возможности реконструкции iCD-SD1.5 при различных конфигурациях. В частности, мы исследуем вклад различных компонентов нашего подхода – числа шагов, функций потерь сохранения и динамического CFG – в качество инверсии. Наша прямая модель работает без CFG ($w = 1$), в то время как для обратной модели мы рассматриваем два режима: ненаправляемый ($w = 1$) и направляемый ($w = 8$), оба из которых важны на практике.

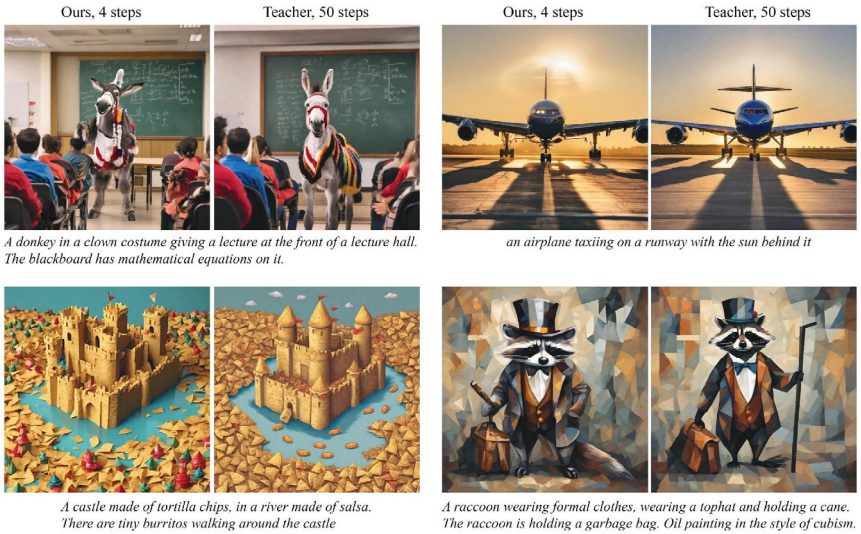

Рис. 6. Несколько примеров генерации изображений по текстовому описанию с помощью модели iCD-XL за 4 шага.

Fig. 6. Few examples of text-to-image generation using the iCD-XL model for 4 steps.

Конфигурация. Для оценки качества инверсии мы используем 5 тыс. изображений и соответствующих текстовых запросов из набора данных MS-COCO [56]. Мы измеряем качество реконструкции с помощью LPIPS [58], PSNR и косинусного расстояния в пространстве признаков DinoV2 [59]. В качестве эталона рассматривается инверсия модели-учителя с отключённым CFG. Для динамического направления мы используем $\tau = 0.7$. Коэффициенты функций потерь сохранения равны $\lambda_f = 1.5$ и $\lambda_r = 1.5$.

Результаты. Результаты представлены в табл. 2. Сначала конфигурации (**A**–**C**) оценивают число шагов вывода прямой и обратной моделей. Мы наблюдаем, что качество реконструкции улучшается с увеличением числа шагов. В наших основных экспериментах мы рассматриваем 3 и 4 шага.

Затем (**E**–**G**) исследуют функции потерь сохранения. В (**E**) мы обучаем прямую модель в режиме энкодера [60], используя только прямую функцию потерь сохранения. Этот эксперимент показывает, что функция потерь согласованности вносит значительный вклад в качество инверсии. (**F**, **G**) показывают, что обе функции потерь улучшают инверсию, причём последняя конфигурация приближается к качеству модели-учителя.

Наконец, мы исследуем динамический CFG и функции потерь сохранения в направляемом режиме декодирования (**I**–**K**) и сравниваем их с настройкой (**H**), не использующей ни одну из этих техник. Из конфигураций (**I**, **J**, **K**) видно, что все техники вносят существенный вклад в качество реконструкции. На рис. 7 мы визуализируем их влияние на инверсию. Видно, что динамический CFG (**I**) отвечает, скорее, за сохранение объектов в целом, тогда как функции

потерь сохранения (**J**, **K**) улучшают, скорее, мелкие детали. Отметим, что итоговая конфигурация (**K**) обеспечивает качество инверсии, сопоставимое с ненаправляемым процессом, при этом сохраняя возможности редактирования благодаря активированному направлению.

Табл. 2. Исследование конфигураций iCD-SD1.5 с точки зрения качества инверсии изображений. Table 2. Exploration of iCD-SD1.5 configurations in terms of image inversion performance.

Ненаправляемый режим декодирования			
Конфигурация	LPIPS ↓	DinoV2 ↑	PSNR ↑
fDDIM^{50} | DDIM^{50}	0.167	0.834	22.98
A fCD^2 | CD^2	0.332	0.632	17.75
B fCD^3 | CD^3	0.317	0.649	18.42
C fCD^4 | CD^4	0.276	0.715	19.19
E ~~fCD^4~~ + L_f | CD^4	0.484	0.554	16.40
F $\text{fCD}^4 + L_f$ | CD^4	0.248	0.728	20.01
G $\text{fCD}^4 + L_f$ | $\text{CD}^4 + L_r$	**0.198**	**0.837**	**22.27**
Направляемый режим декодирования, $w = 8$			
Конфигурация	LPIPS ↓	DinoV2 ↑	PSNR ↑
fDDIM^{50} | DDIM^{50}	0.479	0.534	14.12
fDDIM^{50} | DDIM^{50} + динамический CFG	0.279	0.726	19.58
H fCD^4 | CD^4	0.476	0.550	13.87
I fCD^4 | CD^4 + динамический CFG	0.370	0.650	16.72
J $\text{fCD}^4 + L_f$ | CD^4 + динамический CFG	0.317	0.698	17.98
K $\text{fCD}^4 + L_f$ | $\text{CD}^4 + L_r$ + динамический CFG	**0.273**	**0.749**	**19.66**

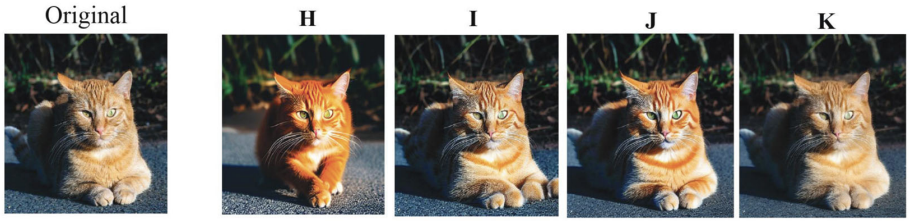

Рис. 7. Влияние динамического направления и функций потерь сохранения на инверсию изображений с помощью iCD.

Fig. 7. Influence of the dynamic guidance and preservation losses on image inversion with iCD.

5.2 Текстово-управляемое редактирование изображений

В этом разделе мы применяем предложенный iCD к задаче редактирования изображений по текстовому описанию. Для модели SD1.5 мы используем метод Prompt-to-Prompt (P2P) [61]. Мы варьируем два гиперпараметра – число шагов cross-attention и self-attention, балансирующих между силой редактирования и сохранением исходного изображения. Мы также применяем наш подход к MasaCTRL [62]. Для модели SDXL мы следуем протоколу оценки ReNoise [25] и лишь изменяем исходный текстовый запрос на этапе декодирования согласно [63].

Метрики. Мы измеряем качество редактирования с помощью автоматических метрик и пользовательского исследования. Первые включают две метрики: 1) для оценки сохранения

исходного изображения мы вычисляем косинусное расстояние между изображениями в пространстве признаков DinoV2; 2) в качестве меры качества редактирования мы используем CLIP score между отредактированным изображением и целевым текстовым запросом. Для пользовательской оценки мы привлекаем профессиональных асессоров, успешно прошедших тестовые задания. Мы показываем им исходный и целевой текстовые запросы, исходное изображение и два изображения, полученных сравниваемыми методами, и задаём вопрос: *какое из отредактированных изображений вы предпочитаете с учётом силы редактирования и сохранения исходного изображения?*

Тестовые наборы. В наших экспериментах мы используем два тестовых набора: PieBench [31] и вручную составленный набор данных COCO на основе пар изображение-текст из MS-COCO [56]. PieBench [31] включает различные типы редактирования, например замену, добавление или удаление объектов. Мы берём 420 примеров реалистичных изображений, охватывающих все типы редактирования. COCO сфокусирован исключительно на задаче замены как одной из наиболее популярных среди практиков. Этот тестовый набор содержит 140 пар изображение-текст.

5.2.1 Текстово-управляемое редактирование с iCD-SD1.5

Конфигурация. Для редактирования с моделью SD1.5 мы используем iCD с 4 прямыми и 4 обратными шагами, обученную с обеими функциями потерь сохранения ($\lambda_f = 1.5, \lambda_r = 1.5$). Мы устанавливаем гиперпараметр динамического CFG $\tau = 0.8$ и максимальный масштаб CFG, равный 19.0.

Базовые методы. Мы сравниваем наш подход с четырьмя базовыми методами, обеспечивающими современное качество редактирования: Null-text Inversion (NTI), Negative-prompt Inversion (NPI) [32], InfEdit [26] и Edit-friendly DDPM [43]. Все методы, кроме InfEdit, являются диффузионными подходами, использующими более 100 шагов. Более того, NTI применяет дополнительную дорогостоящую процедуру оптимизации для повышения качества инверсии. InfEdit работает с дистиллированной диффузионной моделью, latent consistency distillation [18], но использует виртуальную инверсию. Все методы используют P2P, и мы перебираем все возможные значения гиперпараметров, чтобы найти конфигурации, обеспечивающие наилучший компромисс между силой редактирования и сохранением изображения.

Результаты. На рис. 8 и рис. 9 представлены качественные и количественные результаты для обоих тестовых наборов. Мы наблюдаем, что предложенный iCD в большинстве случаев сопоставим с базовыми подходами. Более того, иногда даже превосходит их, будучи при этом в несколько раз быстрее. Например, по результатам пользовательской оценки на наборе COCO наш подход превосходит InfEdit, NPI, NTI и Edit-friendly DDPM (50 шагов). На PieBench он превосходит InfEdit и Edit-friendly DDPM (50 шагов). Кроме того, на рис. 8 мы приводим качественные результаты, подтверждающие конкурентоспособность предложенного метода.

Отметим, что качество предложенного метода несколько ниже на тестовом наборе PieBench по сравнению с COCO как по автоматическим, так и по пользовательским метрикам. Мы связываем это с возросшей сложностью задач редактирования, которая, вероятно, требует большего числа шагов.

5.2.2 Текстово-управляемое редактирование с iCD-XL

Конфигурация. Мы рассматриваем конфигурацию с 4 шагами, $\tau = 0.7$ и масштабом CFG, равным 8.0.

Базовые методы. Мы сравниваем наш подход с недавно предложенным методом ReNoise [25], точно следуя его рекомендациям. Этот метод работает как с дистиллированными моделями (LCM-SDXL [18], SDXL-Turbo [19]), так и с исходной диффузионной моделью

SDXL [1]. Однако даже для дистиллированных моделей требуется значительное число шагов для достижения приемлемого качества.

Результаты. Количественное и качественное сравнения представлены в табл. 3 и на рис. 10 соответственно. По результатам пользовательской оценки iCD-XL превосходит все конфигурации ReNoise. По результатам автоматической оценки наш подход обеспечивает лучшее сохранение исходного изображения (DinoV2 и CLIP score, I), сохраняя при этом высокую способность к редактированию, о чём свидетельствует CLIP score (T).

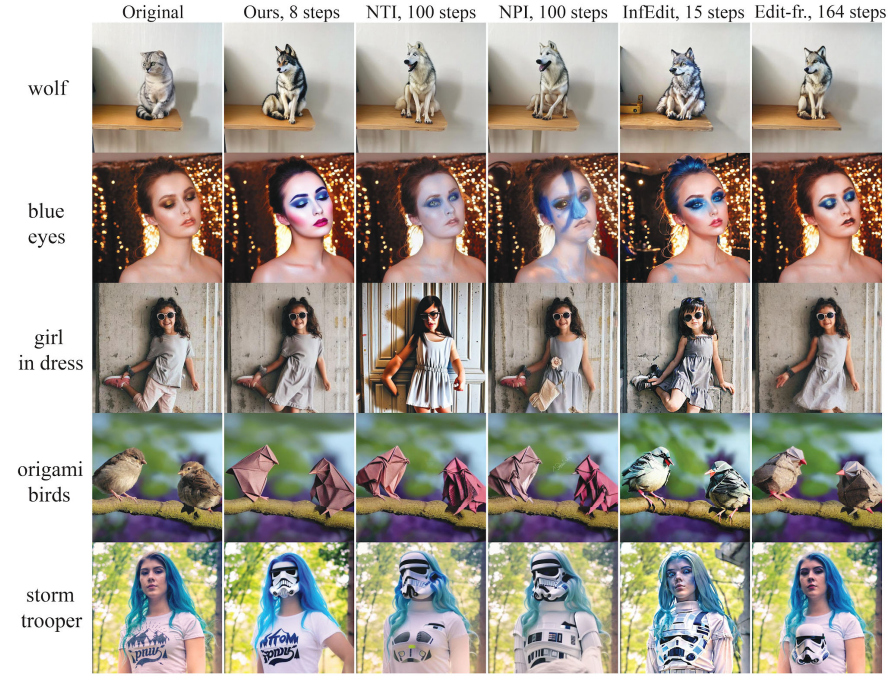

Рис. 8. Примеры редактирования изображений, полученные нашим методом (iCD-SD1.5) и базовыми подходами.
Fig. 8. Image editing examples produced by our method (iCD-SD1.5) and the baseline approaches.

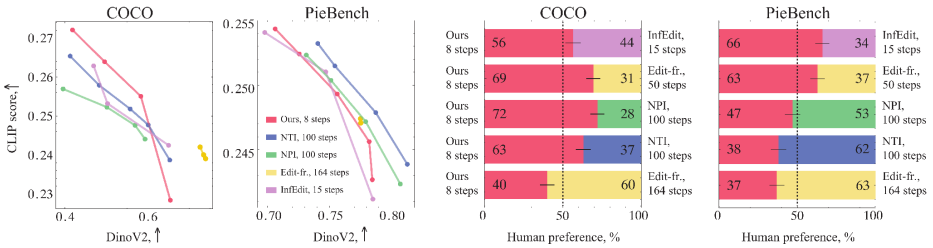

Рис. 9. Количественное сравнение различных подходов к редактированию на основе SD1.5: автоматические метрики (слева) и исследование пользовательских предпочтений (справа).
Fig. 9. Quantitative comparisons between different editing approaches based on SD1.5: automatic metrics (left) and human preference study (right).

Табл. 3. Автоматические метрики (сверху) и оценка пользователями (снизу) для iCD-XL и ReNoise [25].
Table 3. Automatic metrics (top) and human evaluation (bottom) for iCD-XL and ReNoise [25].

Автоматическая оценка, *Тестовый набор COCO*			
Конфигурация	CLIP score, T ↑	DinoV2 ↑	CLIP score, I ↑
iCD, 8 шагов	0.267	**0.472**	**0.748**
ReNoise Turbo, 44 шагов	0.265	0.399	0.705
ReNoise SDXL, 150 шагов	0.227	0.431	0.732
ReNoise LCM, 35 шагов	0.218	0.350	0.663
Автоматическая оценка, *Тестовый набор PieBench*			
Конфигурация	CLIP score, T ↑	DinoV2 ↑	CLIP score, I ↑
iCD, 8 шагов	0.254	**0.707**	**0.858**
ReNoise Turbo, 44 шага	0.254	0.615	0.820
ReNoise SDXL, 150 шагов	0.234	0.642	0.835
ReNoise LCM, 35 шагов	0.236	0.736	0.865

Оценка пользователями, %	
Тестовый набор COCO	*Тестовый набор PieBench*
iCD, 8 шагов: **64 ± 4.9**, ReNoise Turbo, 44 шага: 36 ± 4.9	iCD, 8 шагов: **63 ± 5.1**, ReNoise Turbo, 44 шага: 37 ± 5.1
iCD, 8 шагов: **91 ± 2.8**, ReNoise SDXL, 150: 9 ± 2.8	iCD, 8 шагов: **78 ± 4.3**, ReNoise SDXL, 150: 22 ± 4.3
iCD, 8 шагов: **95 ± 1.8**, ReNoise LCM, 35: 5 ± 1.8	iCD, 8 шагов: **76 ± 4.2,** ReNoise LCM, 35: 24 ± 4.2

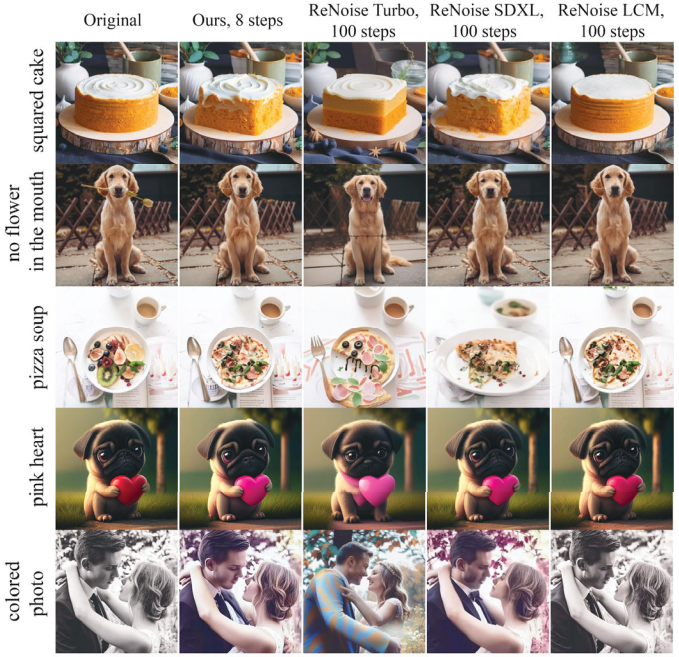

Рис. 10. Примеры редактирования с использованием моделей XL.
Fig. 10. Editing examples using XL models.

6. Заключение

Настоящая статья является русскоязычным переводом и расширенной версией работы, ранее опубликованной на конференции NeurIPS 2024 [64]. В данной работе предложен обобщённый подход дистилляции согласованности, обеспечивающий как быструю инверсию изображений, так и высокое качество генерации за небольшое число шагов семплирования. В сочетании с недавно предложенным динамическим управлением дистиллированные модели демонстрируют высокоэффективные и точные манипуляции с изображениями, что представляет собой значительный шаг к редактированию изображений по текстовому описанию в реальном времени.

Список литературы / References

[1]. D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis.” The Twelfth International Conference on Learning Representations. 2024.

[2]. P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach. “Scaling Rectified Flow Transformers for High-Resolution Image Synthesis.” arXiv preprint arXiv:2403.03206. 2024.

[3]. C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S.K.S. Ghasemipour, R. Gontijo-Lopes, B.K. Ayan, T. Salimans, J. Ho, D.J. Fleet, and M. Norouzi. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.” Advances in Neural Information Processing Systems. 2022.

[4]. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models.” arXiv preprint arXiv:2112.10752. 2021.

[5]. J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, and Z. Li. “PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis.” arXiv preprint arXiv:2310.00426. 2023.

[6]. J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li. “PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation.” arXiv preprint arXiv:2403.04692. 2024.

[7]. T. Brooks, A. Holynski, and A.A. Efros. “Instructpix2pix: Learning to follow image editing instructions.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18392–18402. 2023.

[8]. S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman. “Emu edit: Precise image editing via recognition and generation tasks.” arXiv preprint arXiv:2311.10089. 2023.

[9]. G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu. “Zero-shot image-to-image translation.” ACM SIGGRAPH 2023 Conference Proceedings, pp. 1–11. 2023.

[10]. Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. “Consistency models.” arXiv preprint arXiv:2303.01469. 2023.

[11]. T. Salimans and J. Ho. “Progressive Distillation for Fast Sampling of Diffusion Models.” International Conference on Learning Representations. 2022.

[12]. Y. Song and P. Dhariwal. “Improved techniques for training consistency models.” arXiv preprint arXiv:2310.14189. 2023.

[13]. D. Kim, C.-H. Lai, W.-H. Liao, N. Murata, Y. Takida, T. Uesaka, Y. He, Y. Mitsufuji, and S. Ermon. “Consistency trajectory models: Learning probability flow ode trajectory of diffusion.” arXiv preprint arXiv:2310.02279. 2023.

[14]. W. Luo, T. Hu, S. Zhang, J. Sun, Z. Li, and Z. Zhang. “Diff-Instruct: A Universal Approach for Transferring Knowledge From Pre-trained Diffusion Models.” Thirty-seventh Conference on Neural Information Processing Systems. 2023.

[15]. D. Berthelot, A. Autef, J. Lin, D.A. Yap, S. Zhai, S. Hu, D. Zheng, W. Talbott, and E. Gu. “TRACT: Denoising Diffusion Models with Transitive Closure Time-Distillation.” 2023.

[16]. X. Liu, C. Gong, and Q. Liu. “Flow straight and fast: Learning to generate and transfer data with rectified flow.” arXiv preprint arXiv:2209.03003. 2022.

[17]. L. Li and J. He. “Bidirectional Consistency Models.” arXiv preprint arXiv:2403.18035. 2024.

[18]. S. Luo, Y. Tan, S. Patil, D. Gu, P. von Platen, A. Passos, L. Huang, J. Li, and H. Zhao. “Lcm-lora: A universal stable-diffusion acceleration module.” arXiv preprint arXiv:2311.05556. 2023.

[19]. A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach. “Adversarial Diffusion Distillation.” arXiv preprint arXiv:2311.17042. 2023.

[20]. C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans. “On distillation of guided diffusion models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14297–14306. 2023.

[21]. X. Liu, X. Zhang, J. Ma, J. Peng, and Q. Liu. “Instaflow: One step is enough for high-quality diffusion-based text-to-image generation.” International Conference on Learning Representations. 2024.

[22]. S. Lin, A. Wang, and X. Yang. “SDXL-Lightning: Progressive Adversarial Diffusion Distillation.” arXiv preprint arXiv:2402.13929. 2024.

[23]. A. Sauer, F. Boesel, T. Dockhorn, A. Blattmann, P. Esser, and R. Rombach. “Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation.” arXiv preprint arXiv:2403.12015. 2024.

[24]. T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W.T. Freeman, and T. Park. “One-step Diffusion with Distribution Matching Distillation.” CVPR. 2024.

[25]. D. Garibi, O. Patashnik, A. Voynov, H. Averbuch-Elor, and D. Cohen-Or. “ReNoise: Real Image Inversion Through Iterative Noising.” arXiv preprint arXiv:2403.14602. 2024.

[26]. S. Xu, Y. Huang, J. Pan, Z. Ma, and J. Chai. “Inversion-Free Image Editing with Natural Language.” Conference on Computer Vision and Pattern Recognition 2024. 2024.

[27]. B. Meiri, D. Samuel, N. Darshan, G. Chechik, S. Avidan, and R. Ben-Ari. “Fixed-point Inversion for Text-to-image diffusion models.” arXiv preprint arXiv:2312.12540. 2023.

[28]. J. Ho, A. Jain, and P. Abbeel. “Denoising diffusion probabilistic models.” Advances in neural information processing systems, vol. 33, pp. 6840–6851. 2020.

[29]. Y. Song and S. Ermon. “Generative modeling by estimating gradients of the data distribution.” Advances in neural information processing systems, vol. 32. 2019.

[30]. Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, and B. Poole. “Score-based generative modeling through stochastic differential equations.” arXiv preprint arXiv:2011.13456. 2020.

[31]. X. Ju, A. Zeng, Y. Bian, S. Liu, and Q. Xu. “PnP Inversion: Boosting Diffusion-based Editing with 3 Lines of Code.” International Conference on Learning Representations (ICLR). 2024.

[32]. R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. “Null-text inversion for editing real images using guided diffusion models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6038–6047. 2023.

[33]. D. Miyake, A. Iohara, Y. Saito, and T. Tanaka. “Negative-prompt Inversion: Fast Image Inversion for Editing with Text-guided Diffusion Models.” 2024.

[34]. G. Kim, T. Kwon, and J.C. Ye. “Diffusionclip: Text-guided diffusion models for robust image manipulation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2426–2435. 2022.

[35]. Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu. “Inversion-based style transfer with diffusion models.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10146–10156. 2023.

[36]. J. Ho and T. Salimans. “Classifier-Free Diffusion Guidance.” NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications. 2021.

[37]. B. Wallace, A. Gokul, and N. Naik. “EDICT: Exact Diffusion Inversion via Coupled Transformations.” arXiv preprint arXiv:2211.12446. 2022.

[38]. S. Li, J. van de Weijer, T. Hu, F.S. Khan, Q. Hou, Y. Wang, and J. Yang. “StyleDiffusion: Prompt-Embedding Inversion for Text-Based Editing.” arXiv preprint arXiv:2303.15649. 2023.

[39]. B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani. “Imagic: Text-based real image editing with diffusion models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6007–6017. 2023.

[40]. W. Dong, S. Xue, X. Duan, and S. Han. “Prompt tuning inversion for text-driven image editing using diffusion models.” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7430–7440. 2023.

[41]. D. Miyake, A. Iohara, Y. Saito, and T. Tanaka. “Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models.” arXiv preprint arXiv:2305.16807. 2023.

[42]. L. Han, S. Wen, Q. Chen, Z. Zhang, K. Song, M. Ren, R. Gao, Y. Chen, D. Liu, Q. Zhangli, and others. “Improving negative-prompt inversion via proximal guidance.” arXiv preprint arXiv:2306.05414. 2023.

[43]. I. Huberman-Spiegelglas, V. Kulikov, and T. Michaeli. “An Edit Friendly DDPM Noise Space: Inversion and Manipulations.” arXiv preprint arXiv:2304.06140. 2023.

[44]. M. Brack, F. Friedrich, K. Kornmeier, L. Tsaban, P. Schramowski, K. Kersting, and A. Passos. “LEDITS++: Limitless Image Editing using Text-to-Image Models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024.

[45]. S. Sadat, J. Buhmann, D. Bradley, O. Hilliges, and R.M. Weber. “CADS: Unleashing the Diversity of Diffusion Models through Condition-Annealed Sampling.” The Twelfth International Conference on Learning Representations. 2024.

[46]. T. Kynkäänniemi, M. Aittala, T. Karras, S. Laine, T. Aila, and J. Lehtinen. “Applying Guidance in a Limited Interval Improves Sample and Distribution Quality in Diffusion Models.” arXiv preprint arXiv:2404.07724. 2024.

[47]. A.Q. Nichol and P. Dhariwal. “Improved denoising diffusion probabilistic models.” International conference on machine learning, pp. 8162–8171. 2021.

[48]. T. Karras, M. Aittala, T. Aila, and S. Laine. “Elucidating the design space of diffusion-based generative models.” Advances in Neural Information Processing Systems, vol. 35, pp. 26565–26577. 2022.

[49]. J. Song, C. Meng, and S. Ermon. “Denoising diffusion implicit models.” arXiv preprint arXiv:2010.02502. 2020.

[50]. C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. “Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps.” Advances in Neural Information Processing Systems, vol. 35, pp. 5775–5787. 2022.

[51]. J. Ho and T. Salimans. “Classifier-free diffusion guidance.” arXiv preprint arXiv:2207.12598. 2022.

[52]. S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao. “Latent consistency models: Synthesizing high-resolution images with few-step inference.” arXiv preprint arXiv:2310.04378. 2023.

[53]. J. Heek, E. Hoogeboom, and T. Salimans. “Multistep Consistency Models.” arXiv preprint arXiv:2403.06807. 2024.

[54]. A. Castillo, J. Kohler, J.C. Pérez, J.P. Pérez, A. Pumarola, B. Ghanem, P. Arbeláez, and A. Thabet. “Adaptive Guidance: Training-free Acceleration of Conditional Diffusion Models.” arXiv preprint arXiv:2312.12487. 2023.

[55]. J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. “ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation.” arXiv preprint arXiv:2304.05977. 2023.

[56]. T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, and P. Dollár. “Microsoft COCO: Common Objects in Context.” arXiv preprint arXiv:1405.0312. 2015.

[57]. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. “High-Resolution Image Synthesis With Latent Diffusion Models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684-10695. 2022.

[58]. R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.” CVPR. 2018.

[59]. M. Oquab, T. Darcet, T. Moutakanni, H.V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. “DINOv2: Learning Robust Visual Features without Supervision.” arXiv preprint arXiv:2304.07193. 2023.

[60]. E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. “Encoding in Style: A StyleGAN Encoder for Image-to-Image Translation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2287-2296. 2021.

[61]. A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. “Prompt-to-prompt image editing with cross attention control.” arXiv preprint arXiv:2208.01626. 2022.

[62]. M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng. “Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing.” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22560–22570. 2023.

[63]. C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. “SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations.” International Conference on Learning Representations. 2022.

[64]. N. Starodubcev, M. Khoroshikh, A. Babenko, and D. Baranchuk. “Invertible Consistency Distillation for Text-Guided Image Editing in Around 7 Steps.” *Advances in Neural Information Processing Systems (NeurIPS)*. 2024.

[65]. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).

Информация об авторах / Information about authors

Никита Олегович СТАРОДУБЦЕВ – аспирант Национального исследовательского университета «Высшая школа экономики». Сфера научных интересов: диффузионные модели, дистилляция генеративных моделей, генерация и редактирование изображений по текстовому описанию.

Nikita Olegovich STARODUBCEV – postgraduate student at HSE University. Research interests: diffusion models, generative model distillation, text-to-image generation and editing.