DOI: 10.15514/ISPRAS-2026-38(5)-11

Малошаговые диффузионные модели с прогрессивным увеличением разрешения для генерации изображений и видео

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

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

Аннотация. Современные методы дистилляции диффузионных моделей достигли значительного прогресса, обеспечивая качественную генерацию примерно за 4 шага для крупномасштабных диффузионных моделей генерации изображений и видео по текстовому описанию. Однако дальнейшее уменьшение количества шагов генерации становится всё более сложной задачей. Это наводит на мысль о том, что для повышения эффективности, возможно, стоит обратить внимание на другие степени свободы. Руководствуясь этой идеей, мы представляем SwD – подход для дистилляции диффузионных моделей с постепенным увеличением масштаба генерируемого изображения (Scale-wise Distillation, SwD), который наделяет малошаговые модели способностью к прогрессивной генерации, избегая избыточных вычислений на промежуточных шагах диффузионного процесса. Помимо повышения эффективности, SwD расширяет семейство методов дистилляции, предлагая простую функцию потерь дистилляции на уровне локальных фрагментов изображения или видео, основанную на метрике Maximum Mean Discrepancy (MMD). Эта функция потерь значительно повышает эффективность существующих методов дистилляции и демонстрирует конкурентное качество при использовании в качестве самостоятельного метода. Применённый к современным диффузионным моделям генерации изображений и видео по тексту, SwD приближается по скорости генерации к двум полноразмерным шагам и значительно превосходит альтернативы при сопоставимых вычислительных затратах, что подтверждается автоматическими метриками и исследованиями пользовательских предпочтений.

Ключевые слова: дистилляция диффузионных моделей; генерация изображений и видео по текстовому описанию; малошаговая генерация; вероятностная метрика Maximum Mean Discrepancy.

Для цитирования: Стародубцев Н.О. Малошаговые диффузионные модели с прогрессивным увеличением разрешения для генерации изображений и видео. Труды ИСП РАН, 2026, том 38, вып. 5, стр. 175–194. DOI: 10.15514/ISPRAS-2026-38(5)-11.

Few-step diffusion models with progressive resolution scaling for image and video generation

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. Recent diffusion distillation methods have achieved remarkable progress, enabling high-quality 4-step sampling for large-scale text-conditional image and video diffusion models. However, further reducing the number of sampling steps becomes more and more challenging, suggesting that efficiency gains may be better mined along other model axes. Motivated by this perspective, we introduce SwD, a scale-wise diffusion distillation framework that equips few-step models with progressive generation, avoiding redundant computations at intermediate diffusion timesteps. Beyond efficiency, SwD enriches the family of distribution matching distillation approaches by introducing a simple patch-level distillation objective based on Maximum Mean Discrepancy (MMD). This objective significantly improves the convergence of existing distillation methods and performs surprisingly well in isolation, offering a competitive baseline for diffusion distillation. Applied to state-of-the-art text-to-image/video diffusion models, SwD approaches the sampling speed of two full-resolution steps and largely outperforms alternatives under the same compute budget, as evidenced by automatic metrics and human preference studies.

Keywords: diffusion model distillation; text-to-image and text-to-video generation; few-step generation; Maximum Mean Discrepancy.

For citation: Starodubcev N.O. Few-step diffusion models with progressive resolution scaling for image and video generation. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 5, pp. 175-194 (in Russian). DOI: 10.15514/ISPRAS-2026-38(5)-11.

1. Введение

Разработка крупномасштабных диффузионных моделей (ДМ) для генерации изображений и видео высокого разрешения – по своей сути сложная задача из-за медленной последовательной генерации, требующей 20-50 шагов. Хотя современные ДМ, как правило, работают в скрытом пространстве вариационного автоэнкодера (Variational Autoencoder, VAE) с разрешением примерно в 8 раз ниже исходного [1], скорость генерации остаётся существенным узким местом, особенно для недавних крупномасштабных моделей с более чем 8 миллиардами параметров [2-5]. Эта проблема становится в несколько раз острее для видео-ДМ [6-8], поскольку разрешение скрытого представления масштабируется также и по временному измерению.

Предыдущие работы предпринимали значительные усилия по ускорению ДМ с разных точек зрения [9-13]. Одно из наиболее успешных направлений – дистилляция ДМ в малошаговые генераторы [14-17], где недавние методы [18-19] достигают наилучшего на сегодняшний день качества крупномасштабных диффузионных моделей генерации изображений по тексту примерно за 4 шага. Примечательно, что эти подходы, как правило, сосредоточены на сокращении числа шагов генерации, оставляя неизменными другие перспективные степени свободы, такие как архитектура модели или разрешение входных данных.

Недавно [20-21] провели сравнение между процессом диффузии для изображений и неявной спектральной авторегрессией. В частности, было показано, что обратный диффузионный процесс постепенно предсказывает всё более высокие пространственные частоты, обусловленные ранее сгенерированными низкими частотами через входные данные. Это наблюдение связывает ДМ с моделями предсказания следующего масштаба [22-24], которые также генерируют более высокие пространственные частоты на каждом шаге, но делают это явно, посредством повышения разрешения. Таким образом, результаты наблюдения указывают на значительный потенциал повышения эффективности за счёт выполнения

большинства шагов при более низком разрешении. Однако, несмотря на эту связь, современные малошаговые ДМ по-прежнему работают при фиксированном разрешении на протяжении всего диффузионного процесса.

2. Вклад

- Поскольку большинство современных ДМ относятся к семейству скрытых диффузионных моделей [1], мы прежде всего задаёмся вопросом, применима ли перспектива спектральной авторегрессии также и к скрытым представлениям. В этой работе мы проводим спектральный анализ существующих латентных пространств VAE, а также расширяем его на область видео. Наши результаты подтверждают, что и пространственное, и временное разрешение латентного представления неявно возрастают на протяжении диффузионного процесса, аналогично тому, как это происходит для естественных изображений. Это говорит о том, что латентные ДМ могут избегать избыточных вычислений на промежуточных зашумлённых временных шагах, где высокие частоты в значительной степени подавлены.
- Руководствуясь анализом, мы представляем подход **многомасштабной дистилляции** (Scale-wise Distillation, SwD), который превращает произвольную предобученную ДМ в единую малошаговую модель, постепенно увеличивающую пространственное и временное разрешение объекта на каждом шаге генерации. SwD легко интегрируется с существующими методами дистилляции на основе сопоставления распределений [15-16] и использует их алгоритмы малошаговой генерации, которые оказываются естественным образом согласованными с прогрессивной генерацией.
- Помимо подхода многомасштабной дистилляции, мы представляем простую, но на удивление эффективную целевую функцию дистилляции диффузии, минимизирующую значение метрики Maximum Mean Discrepancy (MMD) [25] в пространстве признаков предобученной ДМ. Предложенная целевая функция дополняет современные методы дистилляции и демонстрирует высокое качество даже в качестве самостоятельного метода, представляя собой конкурентоспособный базовый подход для дистилляции ДМ. Важно, что она не требует дополнительных обучаемых моделей, что делает её вычислительно эффективной и простой для сочетания с существующими подходами дистилляции.
- Мы применяем SwD к современным диффузионным моделям генерации изображений и видео по тексту и показываем, что наши модели конкурируют с моделями-учителями или даже превосходят их, будучи более чем в 10 раз быстрее. По сравнению с полноразмерными малошаговыми моделями, SwD значительно превосходит их при сопоставимых вычислительных затратах. При одинаковом числе шагов генерации SwD обеспечивает примерно двухкратное ускорение при генерации изображений по тексту и примерно трехкратное – при генерации видео по тексту, без потери качества.

3. Обзор литературы

Дистилляция диффузии в малошаговые модели. Методы дистилляции диффузии направлены на сокращение числа шагов генерации до 1-4 при сохранении качества модели-учителя. Эти методы можно в целом разделить на две категории: методы *сопоставления траекторий* (trajectory matching) [26-30], и методы *сопоставления распределений* (distribution matching) [16, 18, 15, 2, 31-33].

Методы сопоставления траекторий аппроксимируют отображение "шум → данные" модели-учителя, интегрируя диффузионное ОДУ за меньшее число шагов, чем численные решатели

[10, 9]. Методы сопоставления распределений ослабляют это ограничение, фокусируясь вместо этого на согласовании распределений студенческой модели и модели-учителя без необходимости точного отображения "шум → данные". Современные подходы, такие как DMD2 [18] и ADD [15, 2], демонстрируют высокое качество генерации всего за 4 шага. Однако они по-прежнему показывают заметное снижение качества при генерации за 1-2 шага, оставляя пространство для дальнейших улучшений. Недавно DMD был успешно адаптирован для видео-диффузионных моделей [19, 34].

Прогрессивная генерация с помощью ДМ. Идея постепенного увеличения разрешения в процессе диффузионной генерации первоначально применялась в иерархических или каскадных ДМ [35-39], которые являются серьёзными конкурентами латентных ДМ [1] для генерации изображений высокого разрешения. Каскадные ДМ состоят из нескольких ДМ, работающих на разных разрешениях, где каждая модель выполняет диффузионную генерацию с нуля, обусловленную представлением более низкого разрешения.

Чтобы связать прогрессивную генерацию с диффузионными процессами, в ряде работ [40-45] были представлены многоэтапные подходы, в которых ДМ обучаются плавно переходить к зашумлённым представлениям более высокого разрешения в процессе диффузионной генерации. SwD продолжает эту линию исследований, предлагая подход, который легко интегрируется в существующие процедуры дистилляции диффузии и адаптирует произвольные предобученные ДМ в прогрессивные малошаговые модели.

Maximum Mean Discrepancy в генеративном моделировании. Maximum Mean Discrepancy (MMD) – это метрика между двумя распределениями P и Q, широко исследованная в ранних работах по GAN [46-50].

Для положительно определённой *функции ядра* $k(x, y)$ MMD может быть определена как

$$MMD^2(P, Q) = E_{x,x' \sim P}[k(x, x')] + E_{y,y' \sim Q}[k(y, y')] - 2\, E_{x \sim P,\, y \sim Q}[k(x, y)], \qquad (1)$$

где x и y обозначают объекты из сгенерированного и целевого распределений соответственно.

Generative Moment Matching Networks (GMMN) [51] используют MMD с фиксированным гауссовским ядром (RBF) непосредственно в пространстве данных. GAN же, напротив, как правило, используют обучаемые ядра, устроенные как композиция дискриминатора с фиксированным ядром.

В диффузионном моделировании MMD применялась для обучения ДМ [52] или их дообучения [53]. DMMD [54] использует адаптированные к шуму дискриминаторы для градиентных потоков MMD [55]. Недавно IMM [56] применила MMD для дистилляции согласованности [14], вычисляя MMD с фиксированным ядром между исходными предсказаниями генератора на разных временных шагах. В нашей работе мы используем MMD между распределениями студенческой модели и модели-учителя в пространстве признаков предобученной ДМ, что даёт мощную и эффективную целевую функцию сопоставления распределений.

4. Спектральный анализ латентного пространства

Работы [20] и [21] показали, что в пространстве пикселей диффузионные модели приближённо реализуют спектральную авторегрессию для естественных изображений. Поскольку современные диффузионные модели генерации изображений по тексту работают с латентными представлениями VAE [1], мы сначала исследуем эту спектральную перспективу для различных латентных пространств, а также распространяем её на область видео.

Следуя [21], мы оцениваем *радиально усреднённую спектральную плотность мощности* (radially averaged power spectral density, RAPSD), то есть усреднённую спектральную

мощность по различным пространственным частотным компонентам, а также её одномерный аналог для временных частот в видео.

Мы исследуем латентные пространства диффузионных моделей изображений и видео, а именно Stable Diffusion 3.5 (SD3.5) [57] и Wan2.1 [7]. VAE модели SD3.5 отображает изображения размером 3×1024×1024 в скрытые представления размером 16×128×128, тогда как VAE модели Wan2.1 кодирует видео размером 81×3×480×832 в представления размером 21×16×60×104. Обе модели используют процесс согласования потоков (flow matching) [58].

Рис. 1 показывает RAPSD гауссовского шума (синий), чистых скрытых представлений (фиолетовый) и зашумлённых представлений (оранжевый) на разных временных шагах. На рис. 1 (слева) показаны результаты для скрытых представлений VAE модели SD3.5. Рис. 1 (справа) показывает среднюю плотность RAPSD как по пространственным, так и по временным частотам скрытых представлений Wan2.1. Вертикальные линии обозначают границы частот: компоненты справа соответствуют высоким частотам, отсутствующим при более низком разрешении, а компоненты слева согласуются с полным разрешением латентного представления (128×128).

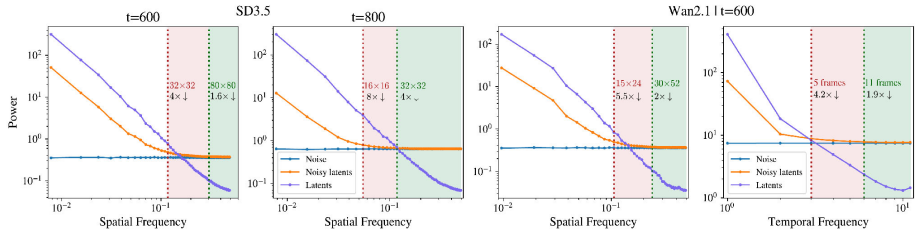

Рис. 1. ***Спектральный анализ*** *скрытых представлений VAE модели SD3.5 (128×128) (слева) и Wan2.1 (21×60×104) по пространственному и временному измерениям (справа). Вертикальные линии отмечают границы частот, справа от которых частотные компоненты отсутствуют в скрытых представлениях более низкого разрешения. Шум маскирует высокие частоты, что говорит о том, что латентные ДМ могут работать при более низком разрешении латентного представления на высоких уровнях шума.* ***Зелёная*** *область показывает допустимое разрешение латентного представления на соответствующем временном шаге, а* ***красная*** *область – что дальнейшее снижение разрешения приводит к заметной потере информации.*

Fig. 1. ***Spectral analysis*** *of SD3.5 VAE latents (128×128) (Left) and Wan2.1 (21×60×104) for spatial and temporal dimensions (Right). Vertical lines mark the frequency boundaries for which the frequency components to the right are not present in lower resolution latents. Noise masks high frequencies, suggesting that latent DMs can operate at lower latent resolutions for high noise levels.*
Green *area indicates the allowed latent resolution at corresponding timestep, while the* ***red*** *area indicates that further resolution reduction leads to noticeable information loss.*

Наблюдения. Во-первых, мы отмечаем, что спектр частот латентного представления приблизительно подчиняется степенному закону, аналогично естественным изображениям [62]. Однако, в отличие от них, наиболее высокочастотные компоненты в латентном пространстве имеют несколько большую амплитуду. Мы связываем это с регуляризационными членами VAE, которые могут приводить к тому, что "чистые" скрытые представления выглядят слегка зашумлёнными.

Мы также наблюдаем, что процесс зашумления постепенно отфильтровывает высокие частоты, тем самым определяя безопасный диапазон понижения разрешения без заметной потери информации. Рис. 1 (слева) показывает, что при $t = 800$ шум маскирует высокочастотные компоненты, возникающие при разрешениях выше 32×32. Это позволяет понизить разрешение скрытых представлений 128×128 в 4 раза (зелёная область). С другой стороны, понижение разрешения в 8 раз исказило бы сигнал данных (красная область), поскольку шум не полностью подавляет эти частоты.

Аналогичный эффект наблюдается и по временному измерению в скрытых представлениях Wan2.1 (см. рис. 1, справа). При $t = 600$ эффективный сигнал может быть представлен примерно 11 латентными кадрами вместо исходных 21.

Практическое значение. Основываясь на этом анализе, мы предполагаем, что латентные диффузионные модели могут работать при более низком разрешении на высоких уровнях шума без потери сигнала данных. Иными словами, моделирование высоких частот на временном шаге t не является необходимым, если эти частоты уже маскированы на данном уровне шума. Отметим, что это справедливо как для пространственной, так и для временной оси в случае видео-ДМ. Мы формулируем этот вывод следующим образом: Латентная диффузия допускает моделирование при более низком разрешении на высоких уровнях шума как по пространственному, так и по временному измерению.

5. Метод

В этом разделе представлен подход **многомасштабной дистилляции** диффузионных моделей, SwD. Сначала мы описываем метод SwD, выделяя его ключевые особенности и трудности. Затем мы представляем нашу целевую функцию дистилляции, основанную на Maximum Mean Discrepancy (MMD).

5.1 Многомасштабная дистилляция диффузионных моделей

Ключевой принцип построения SwD – объединить генерацию на разных масштабах в рамках *единой малошаговой модели* и *единого диффузионного процесса*, в отличие от каскадных подходов. Для этого мы задаём малошаговое *расписание временных шагов*, $[t_1, \dots, t_N]$, и сопоставляем каждому диффузионному временному шагу t_i разрешение латентного представления s_i из *расписания масштабов*, $[s_1, \dots, s_N]$. Таким образом, начиная генерацию с гауссовского шума на наименьшем масштабе s_1, разрешение промежуточных зашумлённых представлений в x_{t_i} постепенно увеличивается по мере выполнения шагов генерации.

Стратегия повышения разрешения. Сначала мы задаёмся вопросом: *как повысить разрешение промежуточного x_{t_i}, чтобы получить достоверный объект более высокого разрешения?* Наивный подход – напрямую повысить разрешение x_{t_i}. Однако это искажает дисперсию шума и вносит локальные корреляции шума. Как следствие, в недавних прогрессивных ДМ [13, 42] предлагаются специальные техники для обработки точек перехода и сохранения непрерывности генерации.

Напротив, малошаговые модели допускают *стохастическую процедуру генерации* [14], предсказывая чистый объект $\hat{x}_0$ и затем повторно зашумляя его до следующего уровня шума. Такой подход естественным образом предполагает повышение разрешения $\hat{x}_0$ перед повторным зашумлением, тем самым сохраняя корректную статистику шума на более высоком разрешении.

Чтобы проверить эту интуицию, мы генерируем изображения с помощью Stable Diffusion 3.5 [57] из промежуточных зашумлённых представлений x_t, полученных с помощью различных стратегий повышения разрешения. А именно, имея скрытое представление реального изображения полного разрешения (128×128), x_0, и его версию с пониженным разрешением (64×64), x_0^{down}, мы рассматриваем: (**A**) эталонную настройку, где шум добавляется к исходному x_0; (**B**) повышение разрешения x_0^{down} с последующим добавлением шума; (**C**) добавление шума к x_0^{down} с последующим повышением разрешения.

Как показано в табл. 1, наивная стратегия (**C**) даёт скрытые представления, в значительной степени выходящие за пределы области данных (OOD). Напротив, стратегия (**B**) показывает результаты, сопоставимые с исходными зашумлёнными представлениями на высоких уровнях шума, вероятно потому, что сильный шум маскирует артефакты интерполяции.

*Табл. 1. Сравнение стратегий повышения разрешения зашумлённых скрытых представлений (**B**, **C**) для перехода 64→128 по качеству генерации (FID-5K). Повышение разрешения x_0^{down} перед добавлением шума (**B**) лучше согласуется с исходными зашумлёнными скрытыми представлениями полного разрешения (**A**).*
*Table 1. Comparison of noisy latent upsampling strategies (**B**, **C**) for 64→128 in generation quality (FID-5K). Upsampling x_0^{down} before noise injection (**B**) aligns better with original full-resolution noisy latents (**A**).*

Конфигурация	$t = 400$	$t = 600$	$t = 800$
A $x_0 \rightarrow^{noise} x_t$	9.2	10.3	13.7
B $x_0^{down} \rightarrow^{upscale} x_0 \rightarrow^{noise} x_t$	30.9	18.8	14.7
C $x_0^{down} \rightarrow^{noise} x_t^{down} \rightarrow^{upscale} x_t$	122	223	327

Мы предлагаем обрабатывать точки перехода между масштабами, повышая разрешение предсказаний $\hat{x}_0$ с последующим повторным зашумлением. На практике мы используем *бикубическую интерполяцию* для пространственных измерений и *смешивание соседних кадров* для временного.

Генерация. SwD выполняет *прогрессивную стохастическую процедуру генерации* в рамках описанного выше подхода к переходу между масштабами. А именно, имея промежуточный зашумлённый объект $\hat{x}_{t_{i-1}}$ на разрешении s_{i-1}, модель предсказывает очищенный объект $\hat{x}_0^{i-1}$. Чтобы перейти к следующему временному шагу t_i, $\hat{x}_0^{i-1}$ повышается в разрешении до s_i и повторно зашумляется прямым диффузионным процессом, в результате чего получается объект более высокого разрешения $\hat{x}_{t_i}$. Затем модель предсказывает следующее $\hat{x}_0^i$. Рис. 2 иллюстрирует этот процесс генерации.

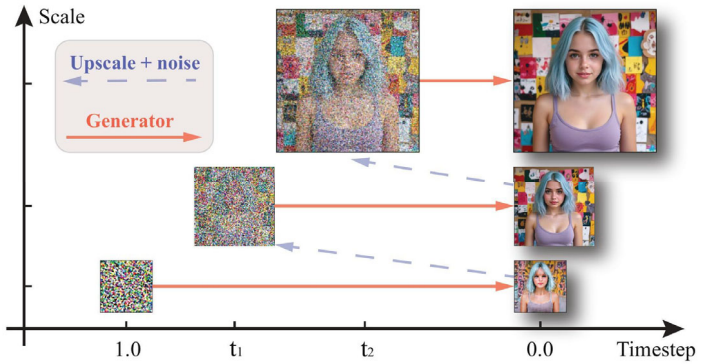

*Рис. 2. **Генерация SwD.** Малошаговая модель начинает генерацию с шума на низком разрешении s_1 и постепенно повышает его в процессе генерации. На каждом шаге предыдущее очищенное предсказание на масштабе s_{i-1} повышается в разрешении и зашумляется в соответствии с расписанием временных шагов, t_i. Затем генератор предсказывает чистое изображение на текущем разрешении s_i.*
*Fig. 2. **SwD sampling.** Few-step model starts sampling from noise at the low resolution s_1 and gradually increases it over generation steps. At each step, the previous denoised prediction at the scale s_{i-1} is upsampled and noised according to the timestep schedule, t_i. Then, the generator predicts a clean image at the current resolution s_i.*

Хотя эта процедура может быть применена непосредственно к предобученным малошаговым моделям, на практике мы наблюдаем, что одного лишь повторного зашумления недостаточно для полного устранения артефактов повышения разрешения, если не используются очень высокие уровни шума. Поэтому мы стремимся обучить малошаговый генератор, который также служит устойчивым инструментом повышения разрешения.

Обучение. Мы обучаем единую модель на нескольких разрешениях, последовательно перебирая пары соседних масштабов $[s_i, s_{i+1}]$ из *расписания масштабов*. На каждом шаге обучения мы берем набор изображений или видео полного разрешения, понижаем их разрешение до исходного и целевого масштабов s_i и s_{i+1} в пространстве пикселей, а затем кодируем их в латентное пространство VAE.

Далее мы повышаем разрешение скрытых представлений более низкого разрешения с s_i до s_{i+1} и применяем прямой диффузионный процесс в соответствии с *расписанием временных шагов*, t_i. Зашумлённые объекты затем подаются на вход многомасштабному генератору, который предсказывает $\hat{x}_0$ на целевом масштабе s_{i+1}.

Мы вычисляем функцию потерь дистилляции между предсказанными и целевыми скрытыми представлениями на s_{i+1}. В нашей работе мы используем сопоставление распределений, вдохновлённое ADD [15, 2] и DMD [16, 18], обеспечивающими наилучшее на сегодняшний день качество дистилляции диффузии. Схематичная иллюстрация этой процедуры обучения представлена на рис. 3.

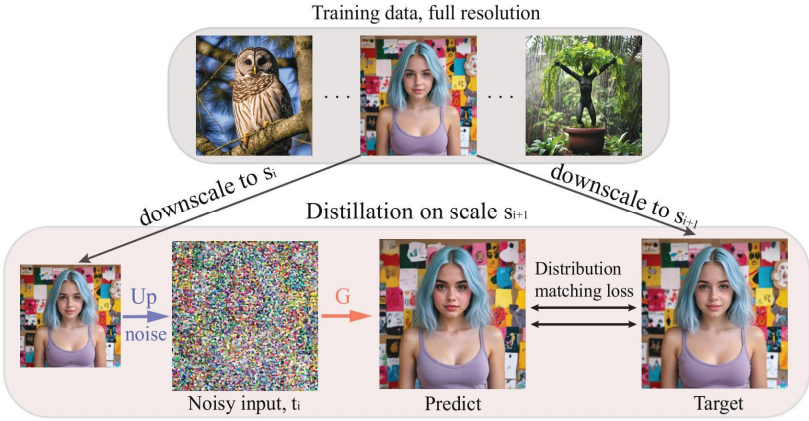

*Рис. 3. **Шаг обучения SwD.** i) Берется пара соседних разрешений $[s_i, s_{i+1}]$ из расписания масштабов. ii) Обучающие изображения понижаются в разрешении до s_i и s_{i+1}. iii) Версии более низкого разрешения повышаются в разрешении и зашумляются до временного шага t_i прямым процессом. iv) По зашумлённым изображениям модель G предсказывает чистые данные на целевом масштабе s_{i+1}. v) Между предсказанными и целевыми изображениями вычисляется функция потерь сопоставления распределений.*
*Fig. 3. **SwD training step.** i) Sample a pair of adjacent resolutions $[s_i, s_{i+1}]$ from the scale schedule. ii) Downscale the training images to s_i and s_{i+1}. iii) The lower scale versions are upsampled and noised to a timestep t_i with the forward process. iv) Given the noised images, the model G predicts clean data at target scale s_{i+1}. v) Distribution matching loss is calculated between predicted and target images.*

Обсуждение расписаний временных шагов и масштабов. Как следует из раздела 4, возникновение высокочастотных компонент при более низких уровнях шума может служить полезным исходным предположением для построения расписаний. Однако, поскольку анализ даёт лишь усреднённые результаты и не учитывает артефакты повышения разрешения, расписания в конечном счёте остаются гиперпараметрами.

На практике мы берём за основу стандартные малошаговые расписания временных шагов и слегка смещаем их в сторону более зашумлённых значений, что согласуется с интуицией о том, что добавление шума помогает сгладить артефакты повышения разрешения. Расписания масштабов монотонно возрастают, начиная с разрешения в 2–4 раза ниже целевого, чтобы добиться желаемого ускорения вывода. Мы также обнаруживаем, что метод не слишком

чувствителен к конкретному выбору расписаний, что позволяет подбирать их с минимальной настройкой и использовать одни и те же расписания для моделей со схожими диффузионными процессами.

5.2 Дистилляция диффузии с использованием Maximum Mean Discrepancy

Помимо предложенного подхода многомасштабной дистилляции, мы расширяем семейство методов дистилляции на основе сопоставления распределений функцией потерь MMD на уровне локальных фрагментов изображений, вычисляемой по промежуточным пространственным признакам предобученных ДМ. Ниже мы обсуждаем вычисление этой функции потерь для трансформерных ДМ [63] – доминирующей архитектуры среди современных ДМ – и отмечаем, что её формулировка без труда обобщается на свёрточные архитектуры.

Во-первых, мы используем способность ДМ работать на разных уровнях шума, что позволяет извлекать структурированные и более детализированные сигналы соответственно на высоких и низких уровнях шума. Соответственно, перед извлечением признаков мы зашумляем как сгенерированные, так и целевые объекты в пределах заданного интервала временных шагов. На практике мы обнаруживаем, что интервал шума от низкого до среднего уровня даёт несколько лучшее качество.

Затем мы извлекаем карты признаков $F \in R^{N\times L\times C}$ из среднего трансформерного блока модели-учителя для сгенерированных и целевых объектов и обозначаем их как F^{fake} и F^{real} соответственно. Здесь N – размер набора данных, L – число пространственных токенов, а C – размерность скрытого представления трансформера. Затем мы вычисляем MMD между распределениями пространственных токенов, которые соответствуют представлениям локальных фрагментов изображения в визуальных трансформерах [64].

Для вычисления MMD мы рассматриваем два ядра: линейное ($k(x,y) = x^T y$) и радиальную базисную функцию (RBF) [65]. Первое согласует средние значения распределений признаков, тогда как второе также согласует моменты более высокого порядка. В наших экспериментах оба ядра показывают схожие результаты, поэтому мы упрощаем L_{MMD}, используя линейное ядро, то есть вычисляем MSE между средними значениями пространственных токенов *для каждого изображения отдельно*:

$$L_{MMD} = \sum_{n=1}^{N} \left\| \frac{1}{L}\sum_{l=1}^{L} F^{real}_{n,l,\cdot} - \frac{1}{L}\sum_{l=1}^{L} F^{fake}_{n,l,\cdot} \right\|^2 . \quad (2)$$

Отметим, что усреднение признаков происходит по всему набору данных, а не по отдельному изображению, как правило, ослабляет специфичную для условия информацию, что в наших экспериментах приводит к более низкому соответствию тексту.

Обсуждение. L_{MMD} с линейным ядром можно рассматривать как адаптацию функции потерь сопоставления признаков (feature matching), предложенной для улучшения обучения GAN [66], применительно к дистилляции диффузии. Насколько нам известно, подобные функции потерь ранее не исследовались в контексте дистилляции диффузии. Ключевые отличия заключаются в следующем: i) L_{MMD} использует предобученную ДМ вместо обучаемого дискриминатора; ii) она использует сигнал с разных уровней шума; iii) средние значения признаков вычисляются для каждого изображения отдельно, а не по всему набору данных.

Итоговая целевая функция. Мы включаем L_{MMD} в качестве дополнительной функции потерь к L_{DMD} и L_{GAN} в нашем многомасштабной подходе: $L_{SwD} = L_{MMD} + \alpha \cdot L_{DMD} + \beta \cdot L_{GAN}$. Интересно, что, несмотря на свою простоту, L_{MMD}, как показано далее, является высококонкурентной самостоятельной целевой функцией дистилляции.

6. Эксперименты

Модели. Мы проверяем наш подход в задаче генерации изображений по тексту, дистиллируя модели SDXL [59], SD3.5 Medium, SD3.5 Large [57] и FLUX.1-dev [3]. Мы также применяем SwD к недавней модели генерации видео по тексту, Wan2.1-1.3B [7].

Данные. Чтобы сохранить постановку изолированной дистилляции и избежать смещений, связанных с внешними данными, мы обучаем все модели исключительно на синтетических данных, сгенерированных моделью-учителем, а не на реальных данных. Отметим, что этот шаг не создаёт узкого места при обучении, поскольку сам процесс дистилляции сходится достаточно быстро (примерно 3 тыс. итераций) и требует значительно меньше данных, чем обучение самой ДМ.

Метрики. Для моделей генерации изображений по тексту мы используем 30 тыс. текстовых запросов из наборов COCO2014 и MJHQ [67, 68] и оцениваем автоматические метрики: FID [69], HPSv3 [70], ImageReward (IR) [71], PickScore (PS) [72] и GenEval [73]. Отметим, что, как было показано, FID слабо коррелирует с человеческим восприятием [72] при оценке генерации изображений по тексту, однако мы приводим эту метрику для полноты картины.

Кроме того, мы проводим исследование пользовательских предпочтений методом попарного сравнения (side-by-side) с участием профессиональных асессоров. Мы отбираем 128 текстовых запросов из набора PartiPrompts [74], следуя [18, 2], и генерируем по 2 изображения на запрос.

Для моделей генерации видео по тексту мы оцениваем VBench-2.0 [75], а также VisionReward [76] и VideoReward [77] на 1003 запросах из набора MovieGenBench [78].

Конфигурация. В наших основных экспериментах мы дистиллируем модели до 4 или 6 шагов. Для моделей генерации изображений по тексту расписания масштабов начинаются с разрешения 256×256 или 512×512 и достигают 1024×1024. Для генерации видео по тексту мы начинаем с разрешения 21×160×272 и достигаем разрешения 81×480×832. Мы выбрали такие начальные точки, поскольку более низкие разрешения дают лишь незначительный прирост скорости.

Также отметим, что SDXL и Wan2.1 дистиллируются с использованием только функции потерь MMD, поскольку функция потерь DMD показывает низкое качество в многомасштабной постановке, когда базовая модель не способна генерировать правдоподобные изображения низкого разрешения.

Бейзлайны. Для генерации изображений по тексту мы сравниваем с моделями-учителями и их общедоступными дистиллированными версиями: DMD2-SDXL [18], SDXL-Turbo [15], Hyper-SD [79], SD3.5-Turbo [2], FLUX-Schnell [3], FLUX-Turbo-Alpha [80]. Мы также сравниваем с другими быстрыми современными моделями, такими как модели предсказания следующего масштаба (Switti [23] и Infinity [24]).

Для задачи генерации видео по тексту мы сравниваем с моделью-учителем (Wan2.1-1.3B [7]) и её 3-шаговым вариантом, дистиллированным методом DMD, – CausVid [19]. Для честного сравнения мы обучаем CausVid с использованием официального кода и настроек, но на нашем обучающем наборе данных, сгенерированном моделью Wan2.1-1.3B, тогда как в публичной версии используются тщательно отобранные внутренние данные.

6.1 Основные результаты

Генерация изображений по тексту. Табл. 2 и рис. 4 показывают сравнение SwD с базовыми методами по качеству генерации и скорости. Результаты сгруппированы по разделам, соответствующим разным базовым диффузионным моделям.

Мы обнаруживаем, что модели SwD демонстрируют наилучшее на сегодняшний день качество по метрикам PS, HPSv3, IR и GenEval в рамках своих семейств моделей. По

задержке вывода SwD показывает почти двухкратное ускорение по сравнению с самыми быстрыми аналогами.

*Табл. 2. Количественное сравнение SwD с другими ведущими моделями с открытым исходным кодом. **Жирным** выделена лучшая модель в каждой группе ДМ, а курсивом – вторая лучшая.*
*Table 2. Quantitative comparison of SwD against other leading open-source models. **Bold** indicates the best-performing model within each DM group, while italic denotes the second best.*

Модель	Шаги	Задержка, с/изобр.	Размер, млрд.	PS↑ COCO	HPSv3↑ COCO	IR↑ COCO	FID↓ COCO	PS↑ MJHQ	HPSv3↑ MJHQ	IR↑ MJHQ	FID↓ MJHQ	GenEval ↑
Switti	14	0.44	2.5	22.6	11.1	0.98	20.0	21.6	9.8	0.84	8.9	0.62
Infinity	14	0.80	2.0	22.7	11.8	0.94	28.1	21.5	10.5	0.98	12.9	0.69
SDXL	50	4.0	2.6	22.5	9.4	0.81	*14.8*	**21.6**	9.2	0.83	**8.1**	0.56
SDXL-Turbo	4	0.12*	2.6	22.6	10.0	0.83	17.5	*21.3*	9.6	0.84	15.4	0.55
Hyper-SD	4	*0.20*	2.6	*22.8*	11.0	*0.90*	20.1	**21.6**	*10.1*	*0.94*	14.7	0.55
SDXL-DMD2	4	*0.20*	2.6	*22.8*	*12.0*	0.87	**14.1**	**21.6**	*10.1*	0.86	*8.3*	**0.58**
SDXL-SwD	4	**0.11**	2.6	**22.9**	**12.4**	**0.95**	21.3	**21.6**	**10.3**	**0.97**	15.1	*0.57*
SD3.5-M	40	4.8	2.2	*22.4*	*10.2*	*1.00*	**16.3**	*21.6*	*9.9*	*0.97*	**9.5**	*0.69*
SD3.5-M-Turbo	4	*0.96*	2.2	22.2	9.6	0.83	*17.6*	21.3	9.3	0.74	13.6	0.59
SD3.5-M-SwD	6	**0.19**	2.2	**22.8**	**11.7**	**1.12**	23.1	**21.8**	**10.7**	**1.10**	*13.4*	**0.70**
SD3.5-L	28	8.3	8.0	**22.8**	*11.3*	*1.06*	**16.5**	**21.8**	*10.4*	*1.04*	**10.7**	*0.70*
SD3.5-L-Turbo	4	*0.63*	8.0	**22.8**	10.0	0.93	22.6	*21.7*	9.9	0.90	*13.5*	*0.70*
SD3.5-L-SwD	4	**0.39**	8.0	**22.8**	**12.8**	**1.20**	*20.6*	**21.8**	**11.1**	**1.22**	13.9	**0.71**
FLUX	30	10.0	12.0	22.9	12.4	1.03	23.6	*21.7*	10.7	0.93	13.0	0.66
FLUX-Turbo-Alpha	8	2.75	12.0	**23.1**	*13.4*	*1.08*	*21.2*	21.5	*11.2*	*0.97*	*11.3*	0.66
Hyper-FLUX	8	2.75	12.0	*23.0*	12.4	0.94	24.2	*21.7*	10.9	0.85	14.9	0.61
FLUX-Schnell	4	*1.41*	12.0	22.6	11.2	1.01	**16.5**	21.5	10.3	0.96	**9.8**	*0.69*
FLUX-SwD	4	**0.72**	12.0	**23.1**	**14.6**	**1.14**	26.4	**21.9**	**11.6**	**1.06**	14.4	**0.71**

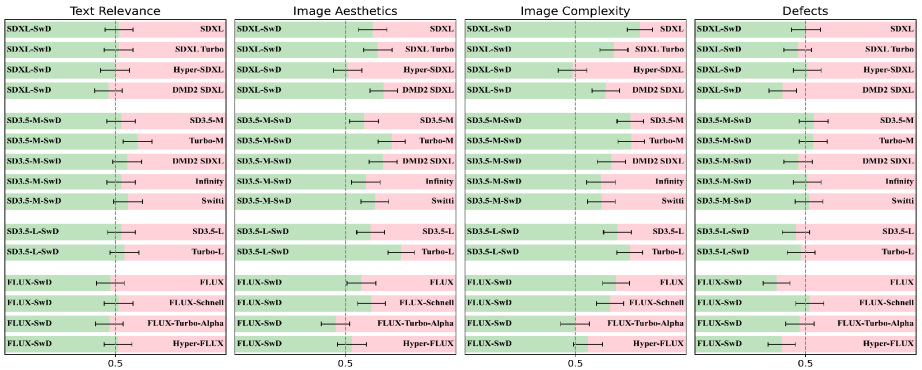

Рис. 4. Исследование пользовательских предпочтений для SwD в сравнении с базовыми моделями.
Fig. 4. Human preference study for SwD against the baseline models.

По результатам пользовательского исследования SwD превосходит большинство других моделей по показателям *сложности изображения* (image complexity) и *эстетичности изображения* (image aesthetics), включая более затратные модели-учителя и их дистиллированные варианты, при сопоставимом уровне *соответствия тексту* (text relevance) и *дефектов* (defects). Мы наблюдаем небольшое ухудшение по дефектам у FLUX-

SwD по сравнению с Hyper-FLUX и FLUX, которые медленнее примерно в 4 и 14 раз соответственно. SDXL-SwD также демонстрирует немного больше дефектов, чем SDXL-DMD2, что мы связываем с использованием только функции потерь MMD, как обсуждается в разделе 6.3. Качественные сравнения представлены на рис. 5.

Генерация видео по тексту. Результаты в табл. 3 показывают, что SwD достигает более высокого качества, чем модель-учитель, будучи при этом в 72 раза быстрее. По сравнению с CausVid SwD показывает сопоставимые результаты, достигая 2.3-кратного ускорения вывода. Мы также обнаруживаем, что SwD, применённая как по временному, так и по пространственному измерению, не ухудшает качество по сравнению с вариантом, использующим только пространственное измерение, и дополнительно повышает скорость вывода. Качественные сравнения представлены на рис. 6.

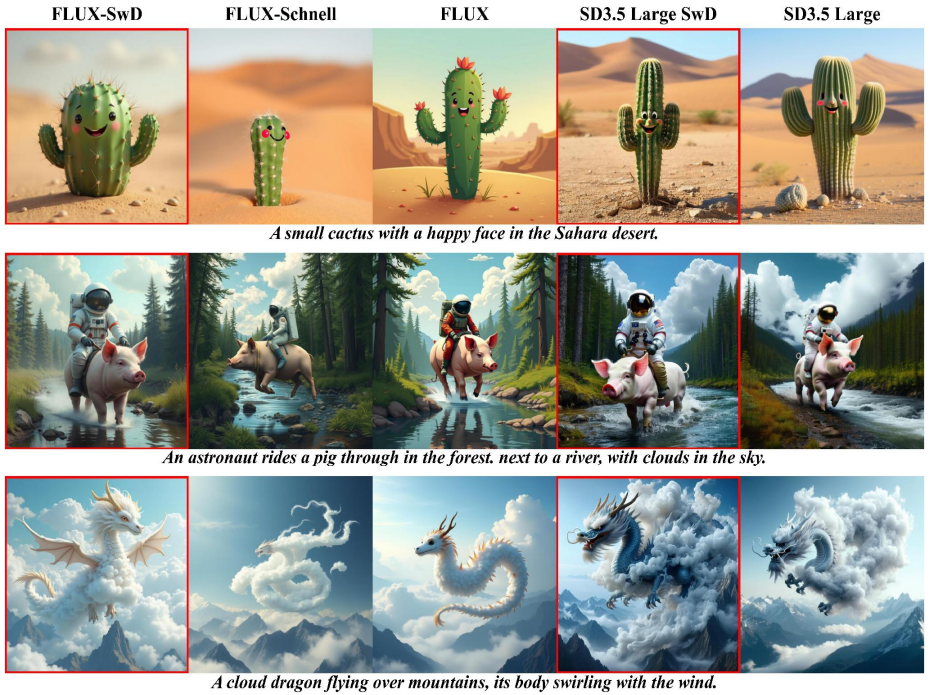

Рис. 5. Качественные результаты для FLUX-SwD и SD3.5 Large SwD.
Fig. 5. Qualitative results of FLUX-SwD and SD3.5 Large SwD.

Табл. 3. Сравнение 4-шагового варианта SwD с моделью Wan 2.1 и 3-шаговой моделью CausVid.
Table 3. Comparison of 4-step SwD variants with the Wan 2.1 and 3-step CausVid models.

Модель	Задержка, с/видео	VisionReward ↑	VideoReward ↑	VBench2 Overall↑
Wan 2.1	137	0.038	5.43	51.6
CausVid*	4.2	0.042	**6.21**	52.3
Spatial SwD	2.1	**0.064**	6.15	52.8
SwD	**1.8**	**0.064**	**6.27**	**53.2**

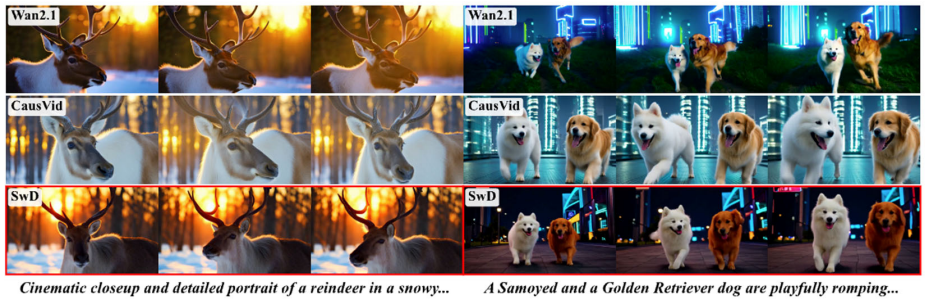

Рис. 6. Качественные результаты для Wan2.1-SwD.
Fig. 6. Qualitative results of Wan2.1-SwD.

6.2 Многомасштабная генерация в сравнении с полноразмерной

Далее мы сравниваем SwD с их полноразмерными аналогами. Полноразмерные базовые модели используют те же расписания временных шагов, но работают с фиксированным целевым разрешением латентного представления.

Мы приводим сравнение качества для SD3.5 Medium. При сравнении настроек с одинаковым числом шагов (4 против 4, 6 против 6) пользовательская оценка не выявляет заметного ухудшения качества. Качественные примеры также подтверждают это. Интересно, что автоматические метрики показывают, что многомасштабные варианты могут даже превосходить свои полноразмерные аналоги, будучи при этом более эффективными.

Затем мы выравниваем время генерации многомасштабных и полноразмерных настроек (4 против 2, 6 против 2 шагов), чтобы оценить различия в качестве. Пользовательская оценка выявляет явное преимущество многомасштабной настройки, особенно в снижении *дефектов* и повышении *сложности изображения*. Примеры на рис. 7 (слева этот рисунок отсутствует, добавлен) показывают высокую частоту дефектов у 2-шаговой полноразмерной базовой модели. Автоматические метрики также стабильно показывают заметный прирост по HPSv3 и PS.

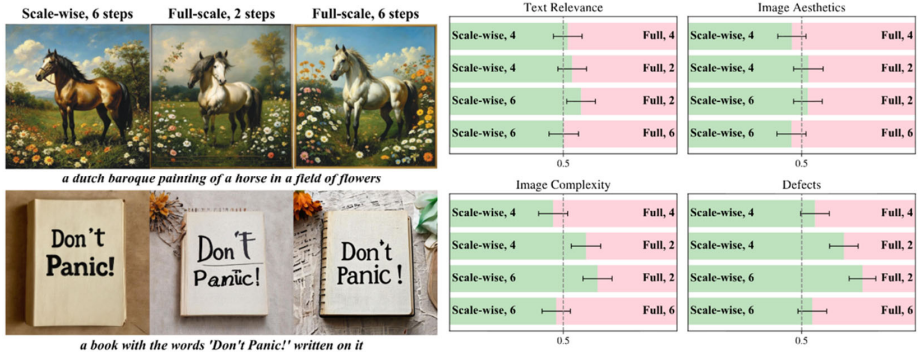

Рис. 7. Наглядные примеры (слева) и исследование пользовательских предпочтений (справа), сравнивающие многомасштабную и полноразмерную конфигурации SD3.5 Medium. Числа указывают количество шагов генерации.
Fig. 7. Visual examples (Left) and human preference study (Right) of the scale-wise and full resolution settings within SD3.5 Medium. The numbers indicate the sampling steps.

Скорость. Табл. 4 приводит задержку генерации одного изображения (включая декодирование VAE и кодирование текста), а табл. 5 показывает среднее время одной итерации обучения. По сравнению с полноразмерной настройкой при том же числе шагов (4 шага) многомасштабная настройка обеспечивает примерно 2-кратное ускорение как обучения, так и генерации для моделей генерации изображений по тексту и примерно 3-кратное – для генерации видео по тексту.

Табл. 4. Время генерации (с/изображение) для многомасштабной и полноразмерной настроек.
Table 4. Sampling times (sec / image) of scale-wise and full-resolution setups.

Настройка	Шаги	SD3.5-M	SD3.5-L	FLUX	Wan2.1
Полноразмерная	4	0.29	0.63	1.41	5.51
Полноразмерная	2	0.16	0.33	0.72	2.97
Многомасштабная	6	0.19	0.41	0.97	2.61
Многомасштабная	4	0.17	0.32	0.72	1.84

Табл. 5. Время обучения (с/итерация) для многомасштабной и полноразмерной 4-шаговых настроек с полной целевой функцией (L_{SwD}) и только с MMD ($L_{SwD-MMD}$).
Table 5. Training times (sec / iteration) for scale-wise and full-resolution 4-step setups using the full objective (L_{SwD}) and MMD only ($L_{SwD-MMD}$).

Настройка	Функция потерь	SD3.5-M	SD3.5-L	FLUX	Wan2.1
Полноразмерная	L_{SwD}	7.5	13.4	22.8	70.6
Полноразмерная	$L_{SwD-MMD}$	1.0	1.7	2.9	12.7
Многомасштабная	L_{SwD}	3.2	7.8	11.3	23.9
Многомасштабная	$L_{SwD-MMD}$	0.4	0.9	1.4	4.4

6.3 Абляция функции потерь MMD

Здесь мы изучаем роль функции потерь L_{MMD} и выбор её конфигурации. Большинство экспериментов проводятся с 6-шаговой настройкой SD3.5-M из раздела 6.1, результаты для MJHQ приведены в табл. 6.

Сначала мы оцениваем вклад L_{MMD} в L_{SwD}. Мы наблюдаем, что обучение только с L_{MMD} немного уступает полной целевой функции L_{SwD}, но остаётся эффективным в качестве самостоятельного метода дистилляции, тогда как исключение L_{MMD} из L_{SwD} приводит к значительному падению качества. Пользовательская оценка (рис. 8) показывает, что модели, обученные только с L_{MMD}, демонстрируют заметное, но не критичное ухудшение по дефектам. Качественные сравнения подтверждают, что они показывают качество, сопоставимое с полной L_{SwD}. Более того, как показано в табл. 5, обучение только с L_{MMD} обеспечивает более чем семикратное ускорение итераций, поскольку не требует обучения дополнительных моделей.

Наконец, мы исследуем несколько вариантов L_{MMD}. L_{MMD} с ядром RBF (**A**) показывает схожие результаты. Опираясь на подход сопоставления признаков [66], мы рассматриваем два изменения: **B** – признаковые токены в формуле (2) усредняются по всему набору данных, а не по отдельному изображению, и **C** – извлечение признаков ДМ только из чистых объектов, без их зашумления диффузионным процессом. Мы наблюдаем, что оба варианта, **B** и **C**, делают L_{MMD} менее эффективной.

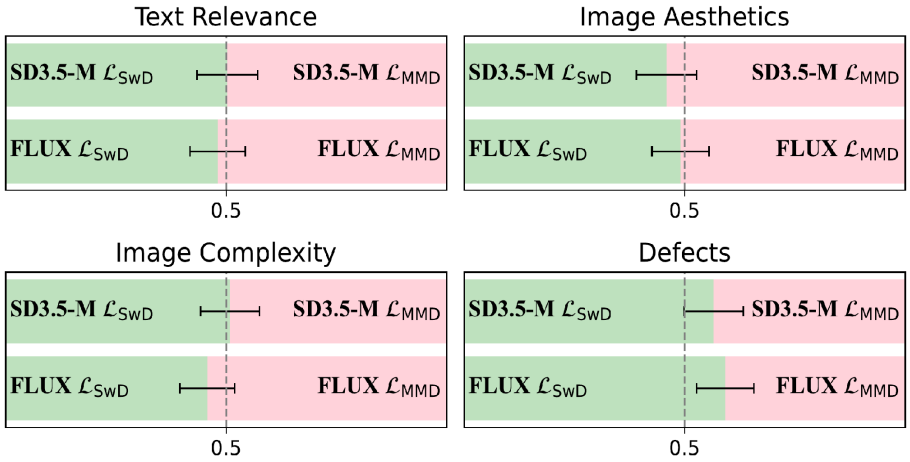

Рис. 8. Сравнение основных моделей SwD с их вариантами, дистиллированными только с помощью L_{MMD}.
Fig. 8. Comparison of the main SwD models against the ones distilled with L_{MMD} alone.

Табл. 6. Абляция целевой функции L_{MMD} для SD3.5-Medium SwD на MJHQ30K.
Table 6. Ablation study of the L_{MMD} objective for SD3.5-Medium SwD on MJHQ30K.

Настройка L_{SwD}	PS↑	HPSv3↑	IR↑	FID↓
L_{SwD} (основная)	**21.8**	**10.7**	1.11	**13.6**
L_{MMD} отдельно	21.5	10.5	**1.15**	13.8
L_{SwD} без L_{MMD}	21.2	9.7	0.91	19.5
A: ядро RBF	**21.8**	**10.8**	1.09	13.7
B: усреднение по набору данных	21.5	10.5	0.97	16.4
C: без зашумления	21.3	10.2	1.01	16.6

6. Заключение

Статья является русскоязычным переводом и расширенной версией работы, принятой на конференции ICLR 2026 [81]. Мы представили SwD – подход многомасштабной дистилляции диффузии, дополненный новой техникой дистилляции на основе MMD на уровне локальных фрагментов изображения или видео. Мы показываем, что обе компоненты могут быть легко объединены с существующими современными методами дистилляции, обеспечивая дальнейшее повышение эффективности и качества малошаговых моделей. Мы полагаем, что предложенная функция потерь для дистилляции ДМ обладает значительным потенциалом для дальнейшего развития на пути к созданию высокоэффективного самодостаточного подхода дистилляции, не требующего дополнительных обучаемых моделей.

Список литературы / References

[1]. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv preprint, 2021. DOI: 10.48550/arXiv.2112.10752.
[2]. Sauer A., Boesel F. et al. Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. arXiv preprint, 2024. DOI: 10.48550/arXiv.2403.12015.
[3]. Black Forest Labs. FLUX.1. Available at: https://huggingface.co/black-forest-labs/FLUX.1-dev.

[4]. Cai Q., Chen J. et al. HiDream-I1: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. arXiv preprint, 2025. DOI: 10.48550/arXiv.2505.22705.
[5]. Wu C. et al. Qwen-Image Technical Report. arXiv preprint, 2025. DOI: 10.48550/arXiv.2508.02324.
[6]. Polyak A. et al. Movie Gen: A Cast of Media Foundation Models. arXiv preprint, 2024. DOI: 10.48550/arXiv.2410.13720.
[7]. Wan T. et al. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv preprint, 2025. DOI: 10.48550/arXiv.2503.20314.
[8]. Kong W., Tian Q., et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint, 2024. DOI: 10.48550/arXiv.2412.03603. 2024.
[9]. Lu C., Zhou Y. et al. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems, 2022, vol. 35, pp. 5775-5787.
[10]. Song J., Meng C., Ermon S. Denoising diffusion implicit models. arXiv preprint, 2022. DOI: 10.48550/arXiv.2010.02502. 2020.
[11]. Wimbauer F., Wu B., Schoenfeld E., Dai X., Hou J., He Z., Sanakoyeu A., Zhang P., Tsai S., Kohler J. et al. Cache me if you can: Accelerating diffusion models through block caching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6211-6220.
[12]. Li X., Liu Y. et al. Q-diffusion: Quantizing diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17535-17545.
[13]. Jin Y., Sun Z et al. Pyramidal Flow Matching for Efficient Video Generative Modeling. The Thirteenth International Conference on Learning Representations, 2025.
[14]. Song Y., Dhariwal P., Chen M., Sutskever I. Consistency models. arXiv preprint, 2023. DOI: 10.48550/arXiv.2303.01469 .
[15]. Sauer A., Lorenz D., Blattmann A., Rombach R. Adversarial Diffusion Distillation. arXiv preprint, 2023. DOI: 10.48550/arXiv.2311.17042.
[16]. Yin T., Gharbi M., Zhang R., Shechtman E., Durand F., Freeman W.T., Park T. One-step Diffusion with Distribution Matching Distillation. CVPR, 2024.
[17]. Kim D., Lai C.-H., Liao W.-H., Takida Y., Murata N., Uesaka T., Mitsufuji Y., Ermon S. PaGoDA: Progressive Growing of a One-Step Generator from a Low-Resolution Diffusion Teacher. arXiv preprint, 2024. DOI: 10.48550/arXiv.2405.14822. 2024.
[18]. Yin T., Gharbi M., Park T., Zhang R., Shechtman E., Durand F., W.T. Freeman. Improved Distribution Matching Distillation for Fast Image Synthesis. NeurIPS, 2024.
[19]. Yin T., Zhang Q., Zhang R., Freeman W.T., Durand F., Shechtman E., Huang X. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. CVPR, 2025.
[20]. Rissanen S., Heinonen M., Solin A. Generative modelling with inverse heat dissipation. International Conference on Learning Representations (ICLR), 2023.
[21]. Dieleman S. Diffusion is spectral autoregression. 2024.
[22]. Tian K., Jiang Y., Yuan Z., Peng B., Wang L. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. arXiv preprint, 2024. DOI: 10.48550/arXiv.2404.02905.
[23]. Voronov A., Kuznedelev D., Khoroshikh M., Khrulkov V., Baranchuk D. Switti: Designing Scale-Wise Transformers for Text-to-Image Synthesis. arXiv preprint, 2024. DOI: 10.48550/arXiv.2412.01819.
[24]. Han J., Liu J., Jiang Y., Yan B., Zhang Y., Yuan Z., Peng B., Liu X. Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. arXiv preprint, 2024. DOI: 10.48550/arXiv.2412.04431.
[25]. Gretton A., Borgwardt K.M., Rasch M.J., Schölkopf B., Smola A. A kernel two-sample test. The journal of machine learning research, 2012, vol. 13, pp. 723-773.
[26]. Meng C., Rombach R., Gao R., Kingma D., Ermon S., Ho J., Salimans T. On distillation of guided diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. pp. 14297-14306.
[27]. Luo S., Tan Y., Huang L., Li J., Zhao H. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference. arXiv preprint, 2023. DOI: 10.48550/arXiv.2310.04378.
[28]. Huang T., Zhang Y., Zheng M., You S., Wang F., Qian C., Xu C. Knowledge Diffusion for Distillation. Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[29]. Song Y., Dhariwal P. Improved Techniques for Training Consistency Models. The Twelfth International Conference on Learning Representations, 2024.

[30]. Kim Y., Anagnostidis S., Du Y., Schönfeld E., Kohler J., Georgopoulos M., Pumarola A., Thabet A., Sanakoyeu A. Autoregressive Distillation of Diffusion Transformers. Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 15745-15756.

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

[32]. Zhou M., Zheng H. et al. Score identity Distillation: Exponentially Fast Distillation of Pretrained Diffusion Models for One-Step Generation. International Conference on Machine Learning, 2024.

[33]. Zhou M., Wang Z., Zheng H., Huang H. Long and Short Guidance in Score identity Distillation for One-Step Text-to-Image Generation. arXiv preprint, 2024. DOI: 10.48550/arXiv.2406.01561.

[34]. Huang X., Li Z., He G., Zhou M., Shechtman E. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. arXiv preprint, 2025. DOI: 10.48550/arXiv.2506.08009.

[35]. Ho J., Saharia C., Chan W., Fleet D.J., Norouzi M., Salimans T. Cascaded Diffusion Models for High Fidelity Image Generation. arXiv preprint, 2021. DOI: 10.48550/arXiv.2106.15282.

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

[37]. Ramesh A., Dhariwal P., Nichol A., Chu C., Chen M. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint, 2022. DOI: 10.48550/arXiv.2204.06125.

[38]. Kastryulin S., Konev A. et al. YaART: Yet Another ART Rendering Technology. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025, vol. 1, pp. 2313-2324.

[39]. Gu J., Zhai S., Zhang Y., Susskind J.M., Jaitly N. Matryoshka diffusion models. The Twelfth International Conference on Learning Representations, 2023.

[40]. Gu J., Zhai S. et al. f-DM: A Multi-stage Diffusion Model via Progressive Signal Transformation. The Eleventh International Conference on Learning Representations, 2023.

[41]. Teng J., Zheng W. et al. Relay Diffusion: Unifying diffusion process across resolutions for image synthesis. arXiv preprint, 2023. DOI: 10.48550/arXiv.2309.03350.

[42]. Atzmon Y., Bala M. et al. Edify Image: High-Quality Image Generation with Pixel Space Laplacian Diffusion Models. arXiv preprint, 2024. DOI: 10.48550/arXiv.2411.07126.

[43]. Zhang Y., Xing J., Xia B., Liu S., Peng B., Tao X., Wan P., Lo E., Jia J. Training-Free Efficient Video Generation via Dynamic Token Carving. arXiv preprint, 2025. DOI: 10.48550/arXiv.2505.16864.

[44]. Anagnostidis S., Bachmann G. et al. Flexidit: Your diffusion transformer can easily generate high-quality samples with less compute. Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 28316-28326.

[45]. Haji-Ali M., Menapace W. et al. Improving Progressive Generation with Decomposable Flow Matching. The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

[46]. Bińkowski M., Sutherland D.J., Arbel M., Gretton A. Demystifying mmd gans." arXiv preprint, 2018. DOI: 10.48550/arXiv.1801.01401.

[47]. Wang W., Sun Y., Halgamuge S. Improving MMD-GAN training with repulsive loss function. arXiv preprint, 2018. DOI: 10.48550/arXiv.1812.09916.

[48]. Dziugaite G.K., Roy D.M., Ghahramani Z. Training generative neural networks via maximum mean discrepancy optimization. arXiv preprint, 2015. DOI: 10.48550/arXiv.1505.03906.

[49]. Bellemare M.G., Danihelka I. et al. The cramer distance as a solution to biased wasserstein gradients. arXiv preprint, 2017. DOI: 10.48550/arXiv.1705.10743.

[50]. Sutherland D.J., Tung H.-Y. et al. Generative models and model criticism via optimized maximum mean discrepancy. arXiv preprint, 2016. DOI: 10.48550/arXiv.1611.04488.

[51]. Li Y., Swersky K., Zemel R. Generative moment matching networks. International conference on machine learning, 2015, pp. 1718-1727.

[52]. Bortoli V.D., Galashov A. et al. Distributional Diffusion Models with Scoring Rules. Forty-second International Conference on Machine Learning, 2025.

[53]. Aiello E., Valsesia D., Magli E. Fast Inference in Denoising Diffusion Models via MMD Finetuning. arXiv preprint, 2023. DOI: 10.48550/arXiv.2301.07969.

[54]. Galashov A., Bortoli V.D., Gretton A. Deep MMD Gradient Flow without adversarial training. The Thirteenth International Conference on Learning Representations, 2025.

[55]. Arbel M., Korba A., Salim A., Gretton A. Maximum Mean Discrepancy Gradient Flow. Advances in Neural Information Processing Systems, 2019, vol. 32.

[56]. Zhou L., Ermon S., Song J. Inductive Moment Matching. Forty-second International Conference on Machine Learning, 2025.

[57]. Esser P., Kulal S. et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. CoRR, 2024, DOI: 10.48550/arXiv.2403.03206.

[58]. Lipman Y., Chen R.T.Q., Ben-Hamu H., Nickel M., Le M. Flow Matching for Generative Modeling. The Eleventh International Conference on Learning Representations, 2023.

[59]. Podell D., English Z. et al. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. The Twelfth International Conference on Learning Representations, 2024.

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

[61]. Song Y., Sohl-Dickstein J., Kingma D.P., Kumar A., Ermon S., Poole B. Score-based generative modeling through stochastic differential equations. arXiv preprint, 2020. DOI: 10.48550/arXiv.2011.13456.

[62]. Schaaf V.D.A., Hateren V.J. Modelling the Power Spectra of Natural Images: Statistics and Information. Vision Research, 1996, vol. 36, pp. 2759-2770.

[63]. Peebles W., Xie S. Scalable Diffusion Models with Transformers. arXiv preprint, 2022. DOI: 10.48550/arXiv.2212.09748.

[64]. Dosovitskiy A., Beyer L. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint, 2020. DOI: 10.48550/arXiv.2010.11929.

[65]. Chang Y.-W., Hsieh C.-J., Chang K.-W., Ringgaard M., Lin C.-J. Training and testing low-degree polynomial data mappings via linear svm. Journal of Machine Learning Research, 2010, vol. 11.

[66]. Salimans T., Goodfellow I., Zaremba W., Cheung V., Radford A., Chen X. Improved Techniques for Training GANs. Advances in Neural Information Processing Systems, 2016, vol. 29.

[67]. Lin T.-Y., Maire M. et al. Microsoft COCO: Common Objects in Context. arXiv preprint, 2015. DOI: 10.48550/arXiv.1405.0312.

[68]. Li D., Kamko A. et al. Playground v2.5: Three Insights towards Enhancing Aesthetic Quality in Text-to-Image Generation. arXiv preprint, 2024. DOI: 10.48550/arXiv.2402.17245.

[69]. Heusel M., Ramsauer H. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 2017, vol. 30.

[70]. Ma Y., Wu X., Sun K., Li H. HPSv3: Towards Wide-Spectrum Human Preference Score. arXiv preprint, 2025. DOI: 10.48550/arXiv.2508.03789.

[71]. Xu J., Liu X. et al. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. Thirty-seventh Conference on Neural Information Processing Systems, 2023.

[72]. Kirstain Y., Polyak A., Singer U., Matiana S., Penna J., Levy O. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation, 2023.

[73]. Ghosh D., Hajishirzi H., Schmidt L. GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment. arXiv preprint, 2023. DOI: 10.48550/arXiv.2310.11513.

[74]. Yu J., Xu Y. et al. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint, 2022. DOI: 10.48550/arXiv.2206.10789.

[75]. Zheng D., Huang Z. et al. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. arXiv preprint, 2025. DOI: 10.48550/arXiv.2503.21755.

[76]. Xu J., Huang Y. et al. VisionReward: Fine-Grained Multi-Dimensional Human Preference Learning for Image and Video Generation. arXiv preprint, 2024. DOI: 10.48550/arXiv.2412.21059.

[77]. Liu J., Liu G. et al. Improving video generation with human feedback. arXiv preprint, 2025. DOI: 10.48550/arXiv.2501.13918.

[78]. Polyak A., Zohar A. et al. Movie Gen: A Cast of Media Foundation Models. arXiv preprint, 2024. DOI: 10.48550/arXiv.2410.13720.

[79]. Ren Y., Xia X. et al. Hyper-SD: Trajectory Segmented Consistency Model for Efficient Image Synthesis. The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.

[80]. AlimamaCreative Team. FLUX.1-Turbo-Alpha. Available at: https://huggingface.co/alimama-creative/FLUX.1-Turbo-Alpha.

[81]. Starodubcev N., Drobyshevskiy I., Kuznedelev D., Babenko A., Baranchuk D. Scale-wise Distillation of Diffusion Models. International Conference on Learning Representations (ICLR), 2026.

Информация об авторах / Information about authors

Стародубцев Никита ОЛЕГОВИЧ – 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.