П. Д. Велижанин, В. В. Ковылина, А. А. Ступников
Тюменский государственный университет, г. Тюмень
УДК 004.89

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АЛГОРИТМОВ ГЕНЕРАЦИИ ВИДЕО НА ОСНОВЕ ДИФФУЗИОННЫХ МОДЕЛЕЙ

Аннотация. В данной работе представлен подход к редактированию реальных видео с использованием предварительно обученных диффузионных моделей без необходимости дополнительного обучения на видеоданных. Для генерации предлагается набор методов, основанных на использовании временной согласованности между кадрами исходного видео для сохранения пространственно-временных связей объектов в кадре генерируемого видео.

Ключевые слова: диффузионные модели, генерация видео, генерация на основе текста, согласованность кадров.

Введение. Диффузионные модели являются разновидностью генеративных моделей и содержат: диффузионный процесс, постепенно добавляющий случайный шум к данным, и процесс декодирования [1], генерирующий новые образцы данных с помощью итеративного удаления шума.

Недавнее успешное применение диффузионных моделей для генерации высококачественных изображений на основе текстового описания [2] мотивировало ряд исследователей к использованию схожей идеи для создания видео [3, 4, 5]. Однако, генерация видео связана со множеством проблем из-за наличия усложненных, по сравнению с изображениями, пространственно-временных связей одних и тех же объектов на разных кадрах [6].

Проблема исследования. В то время, как в области генерации видео на основе текста (TEXT2VID) создаются модели, демонстрирующие все более высокие показатели согласованности [3, 6], генерация на основе видео и текста (TEXT&VID2VID) [4, 5, 7] отстает в результатах. Основное различие заключается в том, что современные TEXT2VID подходы используют технологию 3D UNET [8], позволяющую генерировать видео целиком, при этом TEXT&VID2VID

создает видео покадрово, из-за чего возникает проблема рассогласованности. Под рассогласованностью мы подразумеваем резкое изменение формы и положения объектов в кадре, измеряемое с помощью специальных метрик [9, 10].

Исходя из этого, целью проводимого исследования стали изучение существующих и реализация новых алгоритмов увеличения согласованности TEXT&VID2VID генерации.

Мы создаем новый метод генерации, разделяющий остаточный шум, необходимый для создания изображения, на истинный и связующий, где связующий несет информацию из уже сгенерированных кадров. Кроме того, мы используем новый алгоритм зашумления изображения для увеличения связи между исходным и результирующим видео.

Результаты проведенных экспериментов, представленные ниже, показывают увеличение согласованности на известном видео-датасете при применении описанных алгоритмов как раздельно, так и в комбинации.

Материалы и методы. Диффузионные модели в основном состоят из кодирующей и декодирующей частей. В процессе кодирования, также называемого прямым ходом $q(z_t|x)$, в некоторые данные x цепью Маркова постепенно добавляется случайный шум:

$$z_t = \sqrt{\hat{a}_t}x + \sqrt{1-\hat{a}_t}\epsilon_t, \quad (1)$$

где $\hat{a}_t = \prod_{k=1}^{t} a_k$, $\epsilon_t^i \sim N(0,1)$, a_t — соответствующий коэффициент, от которого зависит, как быстро и насколько много шума добавится в исходные данные. При достаточно больших T, например, $T = 1000$: $\sqrt{\hat{a}_t} \approx 0, \sqrt{1-\hat{a}_t} \approx 1$, а z_t является случайным Гауссовым шумом.

Реализация прямого и обратного процесса зависит от применяемого алгоритма планировщика шума (Noise Scheduler). В рамках данной работы был выбран и далее будет рассматриваться Эйлеров планировщик [11].

Пусть $x = \{x^i | \ i = 1, 2, \dots, N\}$ — исходный видеоклип из N кадров, и $z_t = \{z_t^i \mid i = 1, 2, \dots, N\}$ — зашумленный с помощью прямого процесса кадр x^i на шаге t. Тогда прямой процесс описывается следующим образом:

$$z_t^i = x^i + \sigma_t \epsilon_t^i, \tag{2}$$

где $\epsilon_t^i \sim N(0,1)$, $\sigma_t = \sqrt{\frac{1-\hat{a}_t}{\hat{a}_t}}$.

Для декодирования, или же генерации нового видео $x^* = \{x^{*i} \mid i = 1, 2, \dots, N\}$ из z_t^i применяется обратный процесс:

$$\Delta x_t^i = z_t^{*i} - \sigma_t \, \epsilon_\theta^i\left(z_t^{*i}\right) \tag{3}$$

$$z_{t-1}^{*i} = z_t^{*i} - \frac{(\sigma_t - \sigma_{t-1})\epsilon_\theta^i(z_t^{*i})}{\sigma_t}, \tag{4}$$

где $\epsilon_\theta^i\left(z_t^{*i}\right)$ — вывод сети UNET [8], или же остаточный шум, вычитая который можно получить предсказанное финальное изображение Δx_t^i для шага t.

Данный метод генерации игнорирует связь между кадрами в исходном видео, следовательно, требуется создать новый прямой и обратный процессы, сохраняющие согласованность кадров.

Обратный ход.

Мы описываем новый обратный процесс $p_\theta\left(x \middle| z_t^i, z_t^{i-1}\right)$, разделяющий остаточный шум на истинный $\epsilon_\theta^i\left(z_t^{*i}\right)$, восстанавливающий текущий кадр, и связующий $\mathrm{c}_{\boldsymbol{\theta}}^i\left(z_t^{*i-1}\right)$, восстанавливающий предыдущий кадр. Тогда итоговый остаточный шум записывается следующим образом:

$$r_\theta^i\left(z_t^{*i}\right) = (1 - w)\, \epsilon_\theta^i\left(z_t^{*i}\right) + w \, \mathrm{c}_\theta^i\left(z_t^{*i-1}\right), \tag{5}$$

где $w \in [0, 1]$ является параметром генерации, отражающим степень зависимости z_t^{*i} от z_t^{*i-1}. Так, $w = 0$ означает, что z_t^{*i} не имеет ничего общего с z_t^{*i-1}, но при w = 1 $z_t^{*i} = z_t^{*i-1}$.

Используя предсказанное с помощью остаточного шума предыдущего кадра $\epsilon_\theta^i\left(z_t^{*i-1}\right)$ итоговое изображение Δx_t^{i-1} можно вычислить связующий шум:

$$c_\theta^i\left(z_t^{*i-1}\right) = \left(z_t^{*i} - \Delta x_t^{i-1}\right)\frac{1}{\sigma_t}. \quad (6)$$

В заключительном этапе смешиваются шумы, восстанавливающие изображения Δx_t^{i-1} и Δx_t^i (рис. 1).

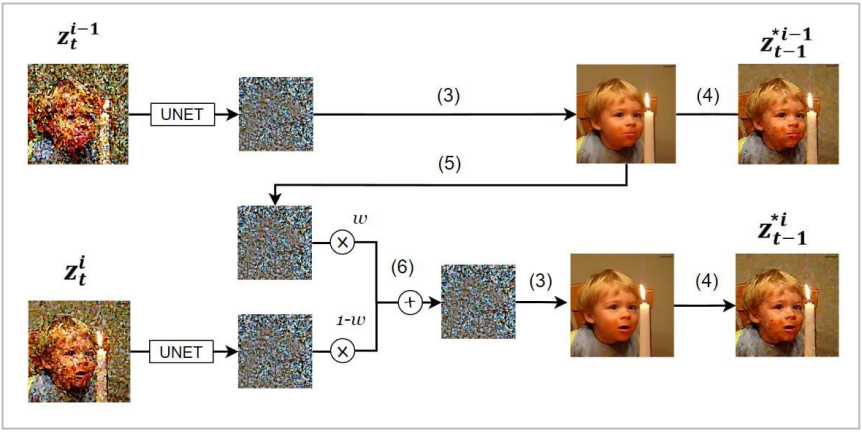

Рис. 1. Смешивание остаточного шума соседних кадров. Переходы обозначены номерами соответствующих уравнений

Прямой ход.

Проблема существующего прямого хода заключается в зависимости алгоритма от добавления случайного Гауссовского шума. Таким образом, при разных $\epsilon_t \sim N(0,1)$ прямой ход $q(z_t|x)$ генерирует различные $z_t = \{z_t(\epsilon_{t,i}) \mid i = 1,2,\dots,N\}$.

Это становится значимым, когда мы применяем обратный ход $p_\theta(x^*|z_t)$, получая $x^* = \{x^{*i} \mid i = 1,2,\dots,N\}$, только приблизительно похожие на исходные данные x. Однако, если мы создадим прямой ход такой, что $x \approx x^*$ (рис. 2), то согласованность видеорефенса будет сохраняться в результирующем видео. В литературе [1, 12] подобные реализации прямого хода называются инверсией.

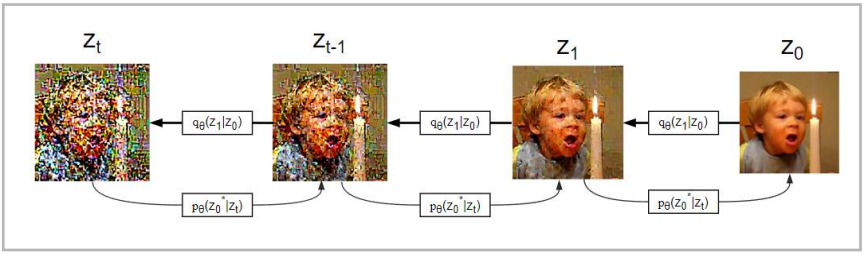

Рис. 2. Визуализация алгоритма инверсии

Первым алгоритмом инверсии является DDIM Inversion [1], а наиболее точным алгоритмом на данный момент можно считать Null-text inversion [12]. Проблема Null-text инверсии заключается в скорости: в среднем для получения релевантного результата длительность работы алгоритма увеличивается в десятки раз, по сравнению с DDIM инверсией.

Нами был реализован алгоритм инверсии, основанный на Эйлеровом планировщике и до этого не применявшийся в генеративных диффузионных моделях. Прямой ход с новой инверсией можно записать следующим образом:

$$z_t = z_{t-1} + (\sigma_t - \sigma_{t-1}) * \frac{\epsilon_\theta(z_{t-1})}{\sigma_t}, \quad (7)$$

где $\epsilon_\theta(z_{t-1})$ — результат работы UNET.

Результаты

Набор данных.

Для оценки созданной модели мы использовали 42 видео из датасета DAVIS [13], содержащих движение людей, машин и животных. Также мы использовали модель BLIP-2 [4] для автоматизации получения текстового описания.

Метрики.

Следуя предыдущим работам [3, 6], для сравнения результатов мы использовали Frechet Video Distance (FVD) [9], как показатель оценки качества сгенерированных моделью клипов. Авторы FVD доказали, что добавление как статического, так и динамического шумов влияет на показатели метрики, поэтому можно утверждать,

что FVD учитывает как визуальное представление объектов видео, так и временную согласованность кадров.

В дополнение, следуя работам, направленных на редактирование видео [4, 5, 7], мы используем метрики:

- CLIP Tem-Con [5] для измерения временной согласованности путем вычисления косинусного сходства между всеми парами последовательных кадров;
- CLIP Frame-Acc [10] для измерения точности редактирования, выражается долей кадров, имеющих более высокое сходство по CLIP с целевым текстовым запросом, чем с исходным;
- PSNR [14] для измерения сходства исходного и сгенерированного видео.

Количественные результаты.

Мы сравниваем результаты с известными методами TEXT& VID2VID редактирования видео на основе TEXT&IMG2IMG диффузионных моделей [4, 5, 7], а также с реализацией редактирования видео с помощью IMG2IMG пайплайна Stable Diffusion [2, 12] и SDEdit [15]. Так, созданная модель повышает согласованность по сравнению с покадровой IMG2IMG генерацией, а результаты ее применения сравнимы с актуальными TEXT&VID2VID методами [4, 7] (табл. 1).

Таблица 1

Сравнение предложенного решения с аналогами для TEXT&VID2VID и TEXT2VID генерации с помощью метрик CLIP Tem-Con, CLIP Fram-Acc, FVD и PSNR на датасете DAVIS.

Метод	*CLIP Tem-Con*↑	*CLIP Fram-Acc*↑	*FVD*↓	*PSNR*↑
Покадровый IMG2IMG (Stable Diffusion) [2]	0.671	0.931	522.41	12.2641
Покадровый IMG2IMG (Stable Diffusion) [2] + Null-text инверсия [12]	0.885	**0.958**	419.95	18.8762
Покадровый SDEdit [15]	0.910	0.819	371.64	**22.9547**
Tune-A-Video [4] + DDIM [1]	0.928	0.750	401.01	9.4568
Наше решение	**0.935**	0.889	**349.53**	19.9512

Качественные результаты.

Мы также проводим визуальное сравнение результатов генерации. На рис. 3 мы демонстрируем сравнение с актуальными методами редактирования видео [4, 7]. Стоит отметить, что, по сравнению с Text2LIVE [7], наше решение может использоваться не только для стилизации, но и изменения формы объектов в кадре. Также, полученные результаты сравнимы с Tune-A-Video [4], при этом, используемой нами TEXT&IMG2IMG модели не требуется дополнительное обучение на входных данных, что в разы ускоряет генерацию. На рис. 4, помимо сохранения согласованности результата, отражено сохранение генеративных способностей полученного метода для различных задач редактирования видео.

Заключение. Генеративные модели обладают практически неограниченными возможностями для творчества, и исследователи в разных предметных областях адаптируют применение подобных сетей для создания изображений, аудио, видео и иного контента.

Наша работа раскрывает возможности генерации видео с помощью модели, изначально обученной для создания изображений. Это может стать мотивацией для дальнейшего исследования потенциала диффузионных моделей обучаться и интерпретировать визуальное представление мира.

В ходе работы создан и реализован метод генерации видео, расширяющий высокоэффективную TEXT&IMG2IMG диффузионную модель [2]. Результаты проведенных экспериментов демонстрируют, что использование информации об уже сгенерированных кадрах и увеличение связи между исходным и результирующим видео может быть эффективным способом генерации видео на основе текста. Хотя созданное решение имеет ограничения: для видео с малым количеством кадров в секунду передаваемой информации оказывается недостаточно, дальнейшие работы могут быть направлены на их устранение.

Рис. 3. Визуальное сравнение созданного решения с Text2LIVE [4] и Tune-A-Video [7]

Изменение атрибутов видео: *squirell, carrot* → ***rabbit, banana***

Изменение стиля видео: ***Van Gogh style***

Рис. 4. Примеры генерации для датасета DAVIS [13]

СПИСОК ЛИТЕРАТУРЫ

1. Jonathan Ho, Ajay Jain, Pieter Abbeel. Denoising diffusion probabilistic models, arXiv:2006.11239v2 [cs.LG] 16 Dec 2020. — Текст : электронный. — С. 1-25. — URL: https://arxiv.org/pdf/2006.11239.pdf (дата обращения: 10.04.2023).
2. High-Resolution Image Synthesis with Latent Diffusion Models arXiv:2112.10752v2 [cs.CV] 13 Apr 2022 / Robin Rombach, Andreas Blattmann, Dominik Lorenz [и др.]. — Текст: электронный. — С. 1-45. — URL: https://arxiv.org/pdf/2112.10752.pdf (дата обращения: 15.01.2023).
3. Video Diffusion Models, arXiv:2204.03458v2 [cs.CV] 22 Jun 2022 / Jonathan Ho, Tim Salimans, Alexey Gritsenko [и др.]. — Текст: электронный. — С. 1-15. — URL: https://arxiv.org/pdf/2204.03458.pdf (дата обращения: 15.01.2023).
4. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation, arXiv:2212.11565v2 [cs.CV] 17 Mar 2023 / Jay Zhangjie Wu, Yixiao Ge, Xintao Wang [и др.]. — Текст: электронный. — С. 1-16. — URL: https://arxiv.org/pdf/2212.11565.pdf (дата обращения: 02.05.2023).
5. Structure and content-guided video synthesis with diffusion models, rXiv:2302.03011v1 [cs.CV] 6 Feb 2023 / Patrick Esser, Johnathan Chiu, Parmida Atighehchian [и др.]. — Текст: электронный. — С. 1-26. — URL: https://arxiv.org/pdf/2302.03011.pdf (дата обращения: 02.05.2023).
6. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation, arXiv:2303.08320v2 [cs.CV] 16 Mar 2023 / Zhengxiong Luo, Dayou Chen, Yingya Zhang [и др.]. — Текст: электронный. — С. 1-10. — URL: https://arxiv.org/pdf/2303.08320v2.pdf (дата обращения: 10.04.2023).
7. Text2LIVE: Text-Driven Layered Image and Video Editing / Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman [и др.]. — Текст: электронный. — С. 1-21. — URL: https://arxiv.org/pdf/2204.02491.pdf (дата обращения: 02.05.2023).
8. Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv:1505.04597v1 [cs.CV] 18 May 2015. — Текст: электронный. — С. 1-8. — URL: https://arxiv.org/pdf/1505.04597.pdf (дата обращения: 10.04.2023).
9. FVD: A NEW METRIC FOR VIDEO GENERATION Published, as a workshop paper at ICLR 2019 / Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach [и др.] — Текст: электронный. — С. 1-9. — URL: https://openreview.net/pdf?id=rylgEULtdN (дата обращения: 15.01.2023).
10. Clipscore: A reference-free evaluation metric for image captioning, arXiv:2104.08718v3 [cs.CV] 23 Mar 2022 / Jack Hessel, Ari Holtzman,

Maxwell Forbes [и др.]. — Текст: электронный. — С. 1-15. — URL: https://arxiv.org/pdf/2104.08718.pdf (дата обращения: 02.05.2023).
11. Elucidating the Design Space of Diffusion-Based Generative Models / Tero Karras, Miika Aittala, Timo Aila, Samuli Laine. — Текст: электронный. — С. 1-21. — URL: https://arxiv.org/pdf/2204.02491.pdf (дата обращения: 02.05.2023).
12. Null-text Inversion for Editing Real Images using Guided Diffusion Models, arXiv:2211.09794v1 [cs.CV] 17 Nov 2022 / Ron Mokady, Amir Hertz, Kfir Aberman [и др.]. — Текст: электронный. — С. 1-20. — URL: https://arxiv.org/pdf/2211.09794.pdf (дата обращения: 10.04.2023).
13. The 2017 DAVIS Challenge on Video Object Segmentation, arXiv:1704.00675v3 [cs.CV] 1 Mar 2018 / Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles [и др.]. — Текст: электронный. — С. 1-6. — URL: https://arxiv.org/pdf/1704.00675.pdf (дата обращения: 15.01.2023).
14. On the Computation of PSNR for a Set of Images or Video, arXiv:2104.14868v1 [eess.IV] 30 Apr 2021 / Onur Keles, M. Akın Yılmaz, A. Murat Tekalp [и др.]. — Текст: электронный. — С. 1-5. — URL: https://arxiv.org/pdf/2104.14868.pdf (дата обращения: 10.04.2023).
15. Sdedit: Image synthesis and editing with stochastic differential equations, arXiv:2108.01073v2 [cs.CV] 5 Jan 2022 / Chenlin Meng, Yang Song, Jiaming Song [и др.]. — Текст: электронный. — С. 1-33. — URL: https://arxiv.org/pdf/2108.01073.pdf (дата обращения: 10.04.2023).