НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2025 Том 25 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
Jule–August 2025 Vol. 25 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2025-25-4-651-662
УДК 004.855.5

Метод генерации анимации цифрового аватара с речевой и невербальной синхронизацией на основе бимодальных данных

Александр Александрович Аксёнов[1]✉, Елена Витальевна Рюмина[2], Дмитрий Александрович Рюмин[3]

[1,2,3] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

[1] axyonov.a@iias.spb.su✉, https://orcid.org/0000-0002-7479-2851
[2] ryumina.e@iias.spb.su, https://orcid.org/0000-0002-4135-6949
[3] ryumin.d@iias.spb.su, https://orcid.org/0000-0002-7935-0569

Аннотация

Введение. Рассмотрена задача генерации анимации цифрового аватара с синхронным воспроизведением речи, мимики и жестикуляции на основе бимодального входа — статического изображения и текста с эмоциональной окраской. Исследована возможность интеграции акустических, визуальных и аффективных признаков в единую модель, обеспечивающую реалистичное и выразительное поведение аватара в соответствии с содержанием и эмоциональным тоном высказывания. **Метод.** Предложенный метод включает шаги извлечения визуальных ориентиров лица, рук и позы, определения пола для выбора подходящего голосового профиля, анализа текста на предмет эмоционального содержания и генерации синтетической аудиоречи. Все признаки интегрируются в генеративной архитектуре на основе диффузионной модели с механизмами временного внимания и межмодального согласования. Это обеспечивает высокую точность синхронизации между речью и невербальными компонентами поведения аватара. Для обучения использовались два специализированных корпуса: один для моделирования жестикуляции, другой — для мимики. Аннотирование производилось средствами автоматического извлечения пространственных ориентиров. **Основные результаты.** Экспериментальное исследование метода выполнялось на многопроцессорной вычислительной платформе с графическими ускорителями. Качество работы модели оценивалось с помощью объективных метрик. Метод показал высокую степень визуального и семантического соответствия: FID — 50,13; FVD — 601,70; SSIM — 0,752; PSNR — 21,997; E-FID — 2,226; Sync-D — 7,003; Sync-C — 6,398. Модель успешно синхронизирует речь с мимикой и жестами, учитывает эмоциональный контекст текста, а также особенности русского жестового языка. **Обсуждение.** Результаты работы могут найти применение в системах эмоционально-чувствительного человеко-машинного взаимодействия, цифровых ассистентах, образовательных и психологических интерфейсах. Предложенный метод представляет интерес для специалистов в области искусственного интеллекта, мультимодальных интерфейсов, компьютерной графики и цифровой психологии.

Ключевые слова

цифровой аватар, BiMoDiCA, мимика, жесты, латентное пространство, генерация анимации, синтез речи, Denoising U-Net, Stable Diffusion

Благодарности

Разделы «Исследования в области генерации цифровых аватаров» и «Экспериментальные исследования метода» выполнены при финансовой поддержке Российского научного фонда (проект № 24-71-00083), остальные исследования — при поддержке того же фонда (проект № 24-71-00112).

Ссылка для цитирования: Аксёнов А.А., Рюмина Е.В., Рюмин Д.А. Метод генерации анимации цифрового аватара с речевой и невербальной синхронизацией на основе бимодальных данных // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 4. С. 651–662. doi: 10.17586/2226-1494-2025-25-4-651-662

A method for generating digital avatar animation with speech and non-verbal synchronization based on bimodal data

Alexander A. Axyonov[1]✉, Elena V. Ryumina[2], Dmitry A. Ryumin[3]

[1,2,3] St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

[1] axyonov.a@iias.spb.su✉, https://orcid.org/0000-0002-7479-2851
[2] ryumina.e@iias.spb.su, https://orcid.org/0000-0002-4135-6949
[3] ryumin.d@iias.spb.su, https://orcid.org/0000-0002-7935-0569

Abstract

This paper addresses the task of generating animations of a digital avatar that synchronously reproduces speech, facial expressions, and gestures based on a bimodal input — namely, a static image and an emotionally colored text. The study explores the integration of acoustic, visual, and affective features into a unified model that enables realistic and expressive avatar behavior aligned with both the semantic content and emotional tone of the utterance. The proposed method includes several stages: extraction of visual landmarks of the face, hands, and body pose; gender recognition for selecting an appropriate voice profile; emotional analysis of the input text; and generation of synthetic speech. All extracted features are integrated within a generative architecture based on a diffusion model enhanced with temporal attention mechanisms and cross-modal alignment strategies. This ensures high-precision synchronization between speech and the avatar nonverbal behavior. The training process utilized two specialized datasets: one focused on gesture modeling, and the other on facial expression synthesis. Annotation was performed using automated spatial landmark extraction tools. Experimental evaluation was conducted on a multiprocessor computing platform with GPU acceleration. The model performance was assessed using a set of objective metrics. The proposed method demonstrated a high degree of visual and semantic coherence: FID — 50.13, FVD — 601.70, SSIM — 0.752, PSNR — 21.997, E-FID — 2.226, Sync-D — 7.003, Sync-C — 6.398. The model effectively synchronizes speech with facial expressions and gestures, accounts for the emotional context of the text, and incorporates features of Russian Sign Language. The proposed approach has potential applications in emotionally aware human — computer interaction systems, digital assistants, educational platforms, and psychological interfaces. The method is of interest to researchers in artificial intelligence, multimodal interfaces, computer graphics, and digital psychology.

Keywords

digital avatar, BiMoDiCA, facial expression, gestures, latent space, animation generation, speech synthesis, Denoising U-Net, Stable Diffusion

Acknowledgements

The “Research on Digital Avatar Generation” and “Experimental Research of the Method” sections were supported by the Russian Science Foundation (Project No. 24-71-00083), while the remaining research was supported by the same Foundation (Project No. 24-71-00112).

For citation: Axyonov A.A., Ryumina E.V., Ryumin D.A. A method for generating digital avatar animation with speech and non-verbal synchronization based on bimodal data. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 4, pp. 651–662 (in Russian). doi: 10.17586/2226-1494-2025-25-4-651-662

Введение

Современные технологии в области искусственного интеллекта позволяют создавать мультимодальные человеко-машинные интерфейсы, которые интегрируют речь, жесты, мимику и эмоции. Однако одним из главных ограничений является нехватка разнообразных мультимодальных корпусов для обучения таких интеллектуальных интерфейсов. В частности, существующие корпуса, охватывающие жестовые языки [1, 2] и распознавание эмоций [3], имеют ограниченный объем и не позволяют создавать универсальные решения для распознавания и синтеза жестов и эмоций. Это затрудняет разработку эффективных технологий для взаимодействия не только людей с ограниченными возможностями, но и для создания более естественных и инклюзивных интерфейсов для всех пользователей.

Одним из перспективных направлений развития мультимодальных интерфейсов является генерация анимации цифрового аватара, способного синхронно воспроизводить речь и сопровождающие ее невербальные сигналы — мимику, жесты и позу. Такая синхронизация способна обеспечить более человеко-машинное взаимодействие, приближая поведение аватара к человеческому. Особенно важным это становится в контексте персонализированных сервисов, где доверие и комфорт играют ключевую роль.

В частности, цифровые аватары с поддержкой речевой и эмоциональной синхронизации могут находить применение в области цифровой психологии и эмоционально-чувствительных систем поддержки и мониторинга психоэмоционального состояния пользователей. Такие аватары потенциально способны выступать в роли виртуальных консультантов или личных психологов, адаптируя свое поведение (интонации, выражение лица, жестикуляцию) в зависимости от психоэмоционального состояния пользователя. Это открывает возможности для создания более эмпатичных и индивидуализированных форм взаимодействия, особенно в условиях ограниченного доступа к квалифицированной помощи или в случаях, когда человеку проще взаимодействовать с цифровым помощником.

Кроме того, подобные технологии могут быть использованы в образовательных и тренировочных системах, где цифровые аватары выступают в роли персональных преподавателей или ассистентов, адапти-

рующих свою невербальную и речевую экспрессию в зависимости от контекста и эмоциональной реакции обучающегося. Это особенно актуально для дистанционного и инклюзивного образования, где высокая степень вовлеченности и индивидуальный подход к обучению играют ключевую роль.

С учетом развития синтетических голосов и генеративных моделей, способных воспроизводить реалистичную артикуляцию и мимику, возрастает интерес к созданию цифровых аватаров. Однако для достижения высокого уровня синхронизации между речью, жестами и эмоциональной экспрессией необходим комплексный метод, способный учитывать как акустические, так и визуальные признаки, включая эмоциональный контекст.

Настоящее исследование направлено на разработку метода генерации анимации цифрового аватара под названием BiModal-based Digital Communication Avatar (BiMoDiCA)[1] с согласованной речевой и невербальной синхронизацией на основе бимодальных данных — изображения и текста с эмоциональной окраской. BiMoDiCA учитывает текущие ограничения в доступных корпусах и демонстрирует потенциал создания более универсальных, гибких и реалистичных цифровых аватаров для широкого спектра задач.

Обзор исследований в области генерации цифровых аватаров

Исследования в области генерации цифровых аватаров охватывают широкий спектр методов, направленных на реалистичную анимацию человека с согласованием речевых и невербальных компонентов[2]. Одним из ключевых направлений является генерация анимации на основе последовательностей поз. В таких методах используется информация о пространственном расположении тела человека, представленная в виде пространственных скелетных ориентиров [4], представлений поз [5], карт глубины [6], трехмерных полигональных сеток [7] или оптического потока [8]. Эти данные служат в качестве управляющих модальностей для генеративных моделей, чаще всего основанных на диффузионных архитектурах [9]. Например, в ряде современных методов применяются модели семейства Stable Diffusion [10, 11] или Stable Video Diffusion [12], в которых информация о позе интегрируется в генеративный процесс посредством модулей управления, аналогичных ControlNet [13]. При этом ключевые точки тела извлекаются с использованием библиотек с открытым исходным кодом, таких как MediaPipe[3] или OpenPose[4], и далее используются в качестве управляющих признаков на этапе подавления шума в генеративной модели.

Параллельно активно развиваются методы, в которых основным источником управляющей информации выступает акустический сигнал. Такие интеллектуальные интерфейсы ориентированы на генерацию движений и мимики, синхронизированных с речью, ее интонационными переходами, акустическими акцентами, ритмом и эмоциональной окраской. Большинство современных исследований сосредоточено на генерации реалистичной анимации лица, в том числе синхронизации движений губ и мимики с аудиоречью. Однако появляются методы, направленные и на генерацию полной жестикуляции. Такие решения демонстрируют высокую степень соответствия между эмоциональными характеристиками речи и визуальными компонентами поведения цифрового аватара. Методы, такие как Vlogger [14], Hallo [15] и MegActor-Sigma [16], используют диффузионные модели для генерации высокодетализированного видеоконтента на основе одного изображения и акустического сигнала, с возможностью управления параметрами мимики и позы. Некоторые методы, такие как CyberHost [17] и EchoMimic [18], реализуют мультимодальное управление и позволяют интегрировать в процессе генерации как акустический сигнал, так и данные о характеристиках выражения лица и телесной позы, включая траектории движений, что обеспечивает более гибкий и точный контроль над синхронизацией речевых и невербальных компонентов в анимации цифрового аватара.

Несмотря на определенный прогресс в области генерации анимации цифрового аватара, существующие методы, как правило, ориентированы либо на управление на основе позы, либо на синхронизацию с акустическим сигналом, редко объединяя обе модальности в единой функциональной схеме. Многие из них ограничиваются областью анимации лица, не учитывая движения тела и жестикуляции, или наоборот – сосредотачиваются на скелетных позах без учета мимики и эмоциональной окраски речи. Кроме того, большинство методов требуют крупномасштабных и специфически размеченных корпусов, что ограничивает их обобщающую способность и применение в реальных условиях. Эти ограничения особенно ощутимы в задачах, где требуется высокая степень согласованности между речью, мимикой и телесной экспрессией, как например, при создании эмпатичных цифровых аватаров.

В данном исследовании предлагается метод генерации анимации цифрового аватара, обеспечивающий речевую и невербальную синхронизации на основе бимодального входа (изображения и текста с эмоциональной окраской). В отличие от существующих решений, данный метод интегрирует визуальные, акустические и эмоциональные признаки в единую диффузионную архитектуру, что обеспечивает более точную и реалистичную синхронизацию анимации лица и тела, а также

1 Название метода предложено авторами настоящей работы.

2 The latest AI-powered technologies usher in a new era of realistic avatars [Электронный ресурс]. Режим доступа: https://huggingface.co/collections/DmitryRyumin/avatars-65df37cdf81fec13d4dbac36, свободный. Яз. англ. (дата обращения: 21.04.2025).

3 MediaPipe [Электронный ресурс]. Режим доступа: https://ai.google.dev/edge/mediapipe/solutions/guide?hl=ru, свободный. Яз. рус. (дата обращения: 21.04.2025).

4 OpenPose [Электронный ресурс]. Режим доступа: https://github.com/CMU-Perceptual-Computing-Lab/openpose, свободный. Яз. англ. (дата обращения: 21.04.2025).

позволяет адаптировать поведение аватара в зависимости от контекста и эмоциональной окраски речи.

Метод

BiMoDiCA представляет собой комплексное решение для генерации видеоконтента, направленный на создание цифровых аватаров, способных воспроизводить реалистичную речевую и жестикуляционную анимацию на основе бимодальных данных, включающих статичное изображение и текст с эмоциональной окраской. Ключевая особенность метода заключается в семантически согласованном объединении вербального и невербального поведений (речевых характеристик, мимики и жестов). Такое объединение обеспечивает высокую степень выразительности создаваемого контента, что критически важно для задач цифровой коммуникации, разработки виртуальных ассистентов, образовательных систем и мультимодальных пользовательских интерфейсов. Функциональная схема предлагаемого метода представлена на рис. 1 и включает ряд компонентов, каждый из которых выполняет специализированную задачу в рамках общей последовательности обработки данных.

На вход метод получает бимодальные данные, включающие статичное изображение человека и короткое текстовое сообщение с выраженной эмоцией. Обозначим изображение как $I \in \mathbb{R}^{H \times W \times 3}$, где H и W — высота и ширина изображения. Изображение человека представляет собой фронтальный портрет, на котором отчетливо видны лицо и верхняя часть тела. Оно служит основой для создания цифрового аватара.

На основе исходного изображения выполняется автоматическое определение пола человека. Изображение I подается на вход предобученной модели Φ_{gender} [19], реализованной в рамках библиотеки с открытым исходным кодом DeepFace[1], которая производит классификацию по бинарному признаку пола и возвращает вектор вероятностей принадлежности к классам «мужчина» и «женщина»:

$$\mathbf{g} = \Phi_{gender}(I) = [g_{male}, g_{female}], \; g_{male}, g_{female} \in [0, 1],$$
$$g_{male} + g_{female} = 1.$$

[1] deepface [Электронный ресурс]. Режим доступа: https://github.com/serengil/deepface, свободный. Яз. англ. (дата обращения: 21.04.2025).

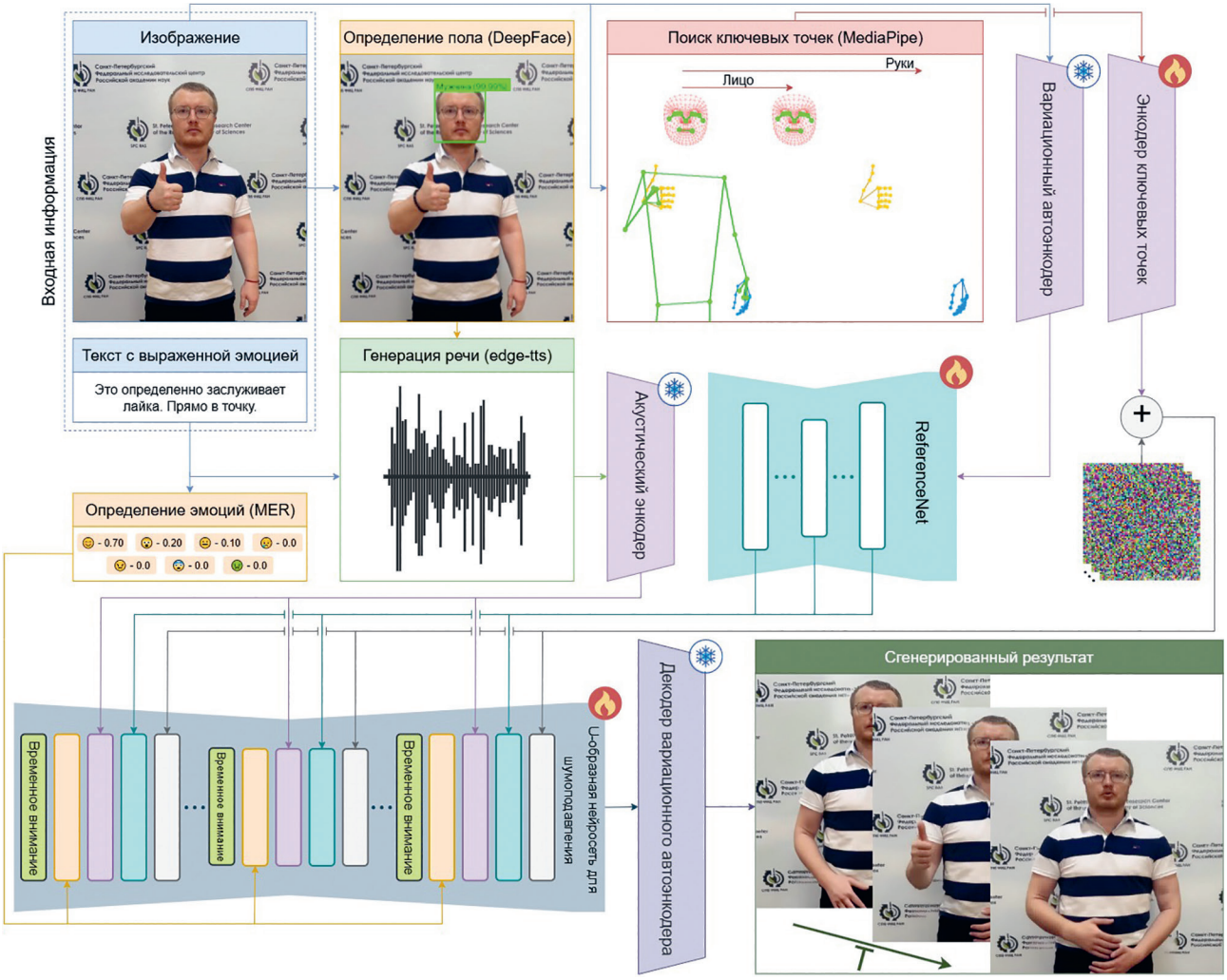

Рис. 1. Функциональная схема предлагаемого метода

Fig. 1. Functional diagram of the proposed method

Таким образом, максимальное значение вектора **g** определяет бинарную метку пола $\hat{y}_{gender} = \text{argmax}(g)$. Значение $\hat{y}_{gender} \in \{male, female\}$ используется для выбора голосового профиля на шаге синтеза речи S, что обеспечивает соответствие между акустической и визуальной модальностями, а также повышает когерентность и реалистичность создаваемого цифрового аватара.

После определения пола, на основе входного изображения I, осуществляется извлечение пространственных ориентиров, необходимых для последующего моделирования артикуляции, мимики и жестов. Для этого используется библиотека с открытым исходным кодом MediaPipe, предоставляющая предобученные модели определения ключевых ориентиров лица [20], рук [21] и тела [22]. Обработка входного изображения I выполняется с помощью предобученной модели[1] $\Phi_{landmarks}$ и возвращает совокупность двувекторного представления позы из 543 двумерных ориентиров, описывающих лицо, руки и тело:

$$\mathbf{L} = \Phi_{landmarks}(I) = \{L^{face}, L^{hands}, L^{pose}\},$$

где $L^{face} = \{(x_i, y_i)\}_{i=1}^{N_f}$ — двумерные координаты $N_f = 468$ ориентиров лица (включая глаза, брови, нос, рот и контур лица); $L^{hands} = \{(x_j, y_j)\}_{j=1}^{2\times N_h}$ — двумерные координаты кистей рук $N_h = 21$ ориентир на каждую руку; $L^{pose} = \{(x_k, y_k)\}_{k=1}^{N_p}$– двумерные координаты $N_p = 33$ ориентиров позы.

Каждая координата ориентира нормирована по размеру исходного изображения I и отражает относительное положение ключевой точки на двумерной сетке. Хотя MediaPipe также возвращает дополнительную компоненту z, описывающую оценку относительной глубины, в рамках предлагаемого метода она не интерпретируется как полноценная пространственная координата, а может использоваться только для эвристической оценки видимости и направленности.

Извлеченные ориентиры **L** представляют собой компактное и согласованное описание визуальной конфигурации, которые подаются в энкодер ключевых точек E_P, генерирующий полную карту позы человека, фиксируя ключевые точки всех частей тела. Эти ориентиры служат базовой позой для цифрового аватара и используются для создания анимации. Однако перед подачей в U-образную нейросеть для шумоподавления (Denoising U-Net) извлеченные ориентиры **L** объединяются с изображениями шума, что помогает улучшить генерацию анимации, за счет более гибкого контроля визуальных характеристик цифрового аватара, а также усилить роль акустического сигнала в управлении артикуляцией и жестами.

Дополнительно исходное изображение I подается на вход вариационного автоэнкодера с дивергенцией Кульбака–Лейблера Φ_{VAE}, реализованного в рамках архитектуры AutoencoderKLMagvit [23], предварительно обученной на крупномасштабных корпусах визуальных анимаций. Несмотря на изначальную ориентацию на обработку видеопоследовательностей, архитектура модели допускает подачу на вход изображения и способна извлекать информативные латентные признаки, отражающие глобальные визуальные характеристики человека. В ходе кодирования осуществляется отображение входного изображения в латентное пространство с последующим получением параметров гауссовского распределения:

$$z_I = \Phi_{VAE}(I) = \mathcal{N}(\mu(I), \sigma^2(I)),$$

где $\mu(I)$ и $\sigma^2(I)$ — параметры латентного распределения, извлекаемые энкодером модели. Извлеченное латентное представление $z_I \in \mathbb{R}^d$, где $d << H \times W$ используется как высокоуровневое описание внешности человека и сохраняется в качестве статического вектора признаков латентного пространства, значительно меньшего по сравнению с размером входного изображения. Такое представление обеспечивает компактное, устойчивое к искажениям кодирование визуального стиля и позволяет сохранить консистентность внешнего вида цифрового аватара на этапах синтеза движений и генерации анимации.

Далее латентное представление z_I, извлеченное исключительно из исходного изображения I, подается в модель усовершенствованной ReferenceNet-ориентированной диффузионной архитектуры [24]. В процессе прохождения через ReferenceNet z_I преобразуется в набор пространственных дескрипторов $f_I = \Phi_{ref}(z_I) \in \mathbb{R}^{C'\times H'\times W'}$, кодирующих высокоуровневые визуальные признаки, такие как форма лица, тела, пальцев рук, прическа, текстура одежды, которые затем внедряются в U-образную нейросеть для шумоподавления. При этом C' — число семантических каналов, каждый кодирует определенный аспект внешнего вида, H' и W' — пространственные размеры (высота и ширина) признакового тензора, которые уменьшаются по сравнению с оригинальным изображением из-за принудительного понижения качества в U-Net. Также важно отметить, что в ReferenceNet не подаются ни акустические признаки, ни представление позы — его задача ограничивается извлечением и сохранением визуального стиля и внешности цифрового аватара.

Второй входной компонент — текстовое сообщение, обозначаемое как $T = \{w_1, w_2, \ldots, w_n\}$, где w_i — отдельное токенизированное слово текста; n — общее количество слов в сообщении. Сообщение представляет собой короткую фразу, включающую как лексическое содержание, так и эмоциональный контекст. Примером может служить утверждение вроде: «Это определенно заслуживает лайка. Прямо в точку.», отражающее позитивную оценку и уверенность говорящего. Текст используется одновременно как основа для генерации синтетической речи и как источник информации об эмоциональном состоянии, которое определяет стиль мимики и жестикуляции в создаваемом видеоконтенте.

Для извлечения эмоциональной составляющей текста применяется адаптированная версия модели из работы [25], сконфигурированная исключительно для текстовой модальности. В исходной архитектуре мето-

[1] MediaPipe Holistic Landmarker [Электронный ресурс]. Режим доступа: https://ai.google.dev/edge/mediapipe/solutions/vision/holistic_landmarker?hl=ru, свободный. Яз. рус. (дата обращения: 21.04.2025).

да [25] реализован механизм перекрестного модального внимания, позволяющий агрегировать признаки из текстовых, акустических и визуальных модальностей. В рамках настоящей задачи входные аудио- и видеоданные отключаются, а механизм внимания перенастраивается так, чтобы все модальное внимание было сосредоточено на текстовом входе с соответствующей нормализацией весов.

На выходе данной модели формируются два типа результатов. Первый — вектор вероятностей по базовым эмоциям:

$$\mathbf{z_E} = [e_1, e_2, \ldots, e_k] \in \mathbb{R}^k, e_i \in [0,1], \sum_{i=1}^{k} e_i = 1,$$

где каждая компонента e_i отражает степень выраженности эмоции $\ell_i \in \mathcal{L}$; $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_k\}$ — фиксированный список базовых эмоций. В рассматриваемой конфигурации используются 7 эмоций:

$$\mathcal{L} = \left\{ \begin{array}{c} \text{счастье, удивление, нейтральность,} \\ \text{злость, грусть, страх, отвращение} \end{array} \right\}.$$

Такой формат представления эмоционального состояния позволяет избежать жесткой категоризации, характерной для классических классификационных моделей, и обеспечивает более гибкое управление невербальным поведением цифрового аватара. Например, результат $\mathbf{z_E}$ = [0,7; 0,2; 0,1; 0; 0; 0; 0] отражает доминирующую эмоцию счастья (с вероятностью 70 %) с оттенками удивления (20 %) и нейтральности (10 %). Такая интерпретация позволяет учитывать не только основную эмоцию, но и вторичные эмоциональные акценты, что повышает достоверность, выразительность и адаптивность поведений цифрового аватара.

Второй тип результата, формируемый на выходе адаптированной модели, представляет собой скрытое векторное представление текста, обобщающее как семантические, так и эмоциональные признаки. Обозначим его как $\mathbf{z_T} \in \mathbb{R}^d$, где d — размерность эмбеддинга. Вектор $\mathbf{z_T}$ не используется напрямую в качестве выходного признака, а применяется в механизмах управления вниманием к мимике и жестикуляции цифрового аватара. Интеграция вероятностного распределения по эмоциям $\mathbf{z_E}$ с латентным текстовым вектором $\mathbf{z_T}$ позволяет сформировать обогащенное представление эмоционально окрашенного текста, которое передается в U-образную нейросеть для шумоподавления. Данное представление сохраняет структурное и смысловое содержания исходной фразы, одновременно обеспечивая эксплицитную привязку к эмоциональному состоянию. На основе этих признаков метод управляет невербальными компонентами поведения (мимикой, движениями головы, руками и позой). Следующий шаг — генерация синтетической аудиоречи и последующая анимация визуальных характеристик.

Для генерации синтетической аудиоречи на основе исходного текста используется современная технология «текст-в-речь» (Text-to-Speech, TTS) — открытая модель[1], реализующая преобразование текстовой последовательности T в аудиосигнал $A \in \mathbb{R}^l$, где l — длина выходной речевой формы. Обозначим данный процесс как $S(T, \hat{y}_{gender}) \rightarrow A$. В настоящем исследовании метод не учитывает эмоциональные параметры при синтезе речи, но учитывает автоматически определенный пол человека. Таким образом, в данном случае TTS работает в нейтральной просодической конфигурации, поскольку основной акцент делается на визуальном выражении эмоций. Тем не менее, добавление эмоциональной модуляции в TTS является перспективным направлением для дальнейших улучшений реалистичности цифрового аватара. Полученный аудиосигнал используется не только как речевой результат, но и как основа для управления артикуляционными компонентами анимации и подается в предобученный акустический энкодер[2], основанный на архитектуре OWSM-CTC [26], параметры которого заморожены на этапе обучения основной модели. Это позволяет использовать акустический энкодер как устойчивый и обобщенный экстрактор признаков. Результатом обработки является векторное представление $\mathbf{z_A} \in \mathbb{R}^m$, где m — размерность аудиопризнакового пространства, равная 1024. Вектор $\mathbf{z_A}$ содержит информацию об артикуляционных особенностях речи, таких как длительность, интонационные переходы и акустические акценты, и используется исключительно в блоках артикуляционной анимации в компоненте U-образной нейросети для шумоподавления — для управления движениями губ, челюсти и нижней части лица.

Ключевым компонентом предлагаемого метода является генеративная архитектура U-образная нейросеть для шумоподавления, построенная на основе модели Stable Diffusion 3.5 Large[3], адаптированной под задачу генерации видеоконтента с цифровым аватаром. Основная задача данного компонента заключается в восстановлении видеокадров из зашумленных представлений с учетом различных условий, поступающих на вход (поза, пространственные дескрипторы, эмоции, текст и аудио). Процесс генерации осуществляется таким образом, что на каждом временном шаге t модель восстанавливает менее зашумленное латентное представление на основе текущего состояния и всех доступных признаков. Этот процесс можно описать выражением:

$$x_{t-1} = D_\theta(\mathbf{L}, f_l, \mathbf{z_E}, \mathbf{z_T}, \mathbf{z_A}, x_t, t),$$

где x_t — зашумленное латентное представление на шаге t; D_θ — U-образная нейросеть для шумоподавления с параметрами θ.

Модифицированная U-образная нейросеть для шумоподавления организована в виде энкодер-декодера, сохраняя L-уровневую U-образную схему. Энкодер использует последовательность сверточных блоков с операциями понижения разрешения (down-sampling),

[1] edge-tts [Электронный ресурс]. Режим доступа: https://github.com/rany2/edge-tts, свободный. Яз. англ. (дата обращения: 21.04.2025).

[2] owsm_ctc_v3.2_ft_1B [Электронный ресурс]. Режим доступа: https://huggingface.co/espnet/owsm_ctc_v3.2_ft_1B, свободный. Яз. англ. (дата обращения: 21.04.2025).

[3] Stable Diffusion 3.5 Large [Электронный ресурс]. Режим доступа: https://huggingface.co/stabilityai/stable-diffusion-3.5-large, свободный. Яз. англ. (дата обращения: 21.04.2025).

извлекая иерархические пространственные признаки, а декодер восстанавливает детали через операции повышения разрешения (up-sampling) и skip-соединения, которые передают высокочастотные детали из энкодера, критичные для сохранения четкости анимации. Взаимодействие признаков выполняется иерархически. Во-первых, векторные представления позы $\mathbf{L}$ и пространственных дескрипторов f_l подаются на слой линейной модуляции, обеспечивающий адаптивную нормализацию признаков. Во-вторых, вводятся эмоциональные векторы $\mathbf{z_E}$ и текстовый эмбеддинг $\mathbf{z_T}$ исключительно в узловую часть сети, фокусируя наиболее компактное представление на семантике высказывания. Затем, интегрируются аудиопризнаки $\mathbf{z_A}$ на этапах up-sampling, уточняя артикуляцию губ и нижней части лица. Такое разграничение потоков позволяет изолированно контролировать позу, мимику и артикуляцию, снижая взаимные помехи модальностей и улучшая синхронизацию.

Для повышения временной согласованности между соседними кадрами в архитектуру встроены модули временного внимания, которые моделируют краткосрочные зависимости внутри видеопоследовательности. Модуль вычисляет веса взаимодействия $softmax(QK^T)V$ между кадрами через механизм на основе скалярного произведения [27]. На вход подаются линейно спроецированные матрицы запроса Q, ключей K и значений V для исходных признаков. Механизм фокусирует внимание на областях с артефактами, такими как дрожание головы, рук или скачки мимики. Это позволяет устранить визуальные артефакты за счет учета контекста предыдущих и последующих кадров.

Дополнительно реализован механизм согласования модальностей, обеспечивающий совместную обработку информации из различных модальностей (поза, пространственные дескрипторы, эмоции, текст и аудиосигнал). В процессе внимания пространственные дескрипторы используются в роли запросов, а признаки из других модальностей — в роли ключей и значений. Их взаимодействие агрегируется с исходными признаками через остаточное соединение, что позволяет комплексировать разные модальности без потери их информативности.

Благодаря использованию архитектуры Stable Diffusion 3.5 Large в качестве генеративного ядра обеспечивается высокая визуальная детализация, реалистичность текстур лица и тела, а также стабильность визуального потока между кадрами. Выходом является последовательность видеокадров, в которых движения цифрового аватара точно соответствуют содержанию и интонации сгенерированной аудиоречи.

Экспериментальное исследование метода

В ходе проведенных исследований была выполнена комплексная оценка предлагаемого метода генерации анимаций цифровых аватаров, использующего изображение и эмоционально окрашенный текст в качестве входной информации. Оценка охватывала несколько ключевых аспектов, включая визуальное качество сгенерированного видеоконтента, точность синхронизации мимики и речи, а также способность модели достоверно передавать эмоции через жестикуляцию. Исследованы различные конфигурации архитектур и значений гиперпараметров, что позволило определить наиболее значимые компоненты метода для достижения визуально и семантически реалистичных результатов. Кроме того, внимание уделено синхронизации движений тела и жестов в контексте русского жестового языка, что позволяет моделировать коммуникацию как для людей с нарушениями слуха, так и для обычных пользователей. Метод также может генерировать синтетические данные для обучения систем распознавания жестов.

Аппаратные и программные настройки. Все эксперименты выполнялись с использованием высокопроизводительной вычислительной инфраструктуры на базе графических ускорителей NVIDIA A100 с 80 ГБ видеопамяти. Обучение велось в распределенном режиме на 8 GPU, что обеспечивало эффективное масштабирование и стабильную сходимость даже для ресурсоемких моделей, включая ReferenceNet и U-образную нейросеть для шумоподавления. Программная среда была построена на версии Python 3,12 с использованием библиотеки PyTorch 2,4 и поддержки архитектуры Compute Unified Device Architecture. В качестве операционной системы применялась CentOS 7 с ядром Linux, обеспечивающая стабильную и воспроизводимую среду для проведения всех экспериментов. Управление зависимостями и конфигурацией окружения осуществлялось через платформу Docker с зафиксированными начальными значениями генерации (seed).

Используемые корпуса и аннотирование. Для обучения необходимых нейросетевых моделей были задействованы два корпуса, охватывающие разные аспекты визуальной выразительности. Для моделирования жестикуляций применялся корпус русского жестового языка Slovo [28], содержащий 20 400 видео с жестами от 194 участников, представленных в разнообразных контекстах. Для генерации мимики применен корпус VFHQ [29], включающий более 16 000 видео с детализированными лицевыми выражениями в условиях интервью. Оба корпуса были аннотированы с использованием MediaPipe, что обеспечило точное извлечение ключевых ориентиров лица, рук и тела, способствуя синхронизации визуальных компонентов с речью и повышая реалистичность поведения цифровых аватаров. При аннотировании жестов и мимики учитывались особенности русского жестового языка [30, 31], что повысило точность синхронизации.

Параметры обучения и стратегия оптимизации. Обучение компонентов предложенного метода проводилось с использованием тщательно подобранных гиперпараметров и современных стратегий оптимизации, которые были выбраны на основе серии экспериментов для достижения наилучших результатов в плане стабильности сходимости и обобщающей способности всех обучаемых нейросетевых моделей.

Для оптимизации каждой модели использовался адаптивный оптимизатор AdamW с коэффициентом веса регуляризации равным 0,01, что помогло избежать переобучения. Начальная скорость обучения была

установлена на уровне $1 \cdot 10^{-4}$ и адаптивно изменялась с использованием косинусного отжига [32].

При этом первые 5 % шагов (около 20 000 итераций) проходили с фазой линейного разогрева [33], что позволило избежать резкого старта и обеспечило более плавную настройку параметров модели на первых шагах обучения. На финальных этапах обучения применялась техника стохастического усреднения весов [34] с шагом 500 итераций, что также способствовало снижению колебаний и улучшению стабильности предсказаний между эпохами.

В качестве функции потерь использовалась комбинация нескольких методов, адаптированных под специфику каждого компонента метода. Для ReferenceNet была применена комбинированная функция потерь, включающая среднеквадратичную ошибку, и перцептивную потерю согласованности кадров [35], которая измеряла сходство между соседними кадрами, что особенно важно для устойчивости динамических анимаций.

Энкодер ключевых точек был обучен с использованием Mean Absolute Error. Для повышения устойчивости к отсутствующим данным случайным образом исключались 15 % ключевых точек. Это позволило улучшить зависимость модели от других источников данных, таких как акустические и эмоциональные признаки.

После завершения обучения ReferenceNet и энкодера ключевых точек эти компоненты использовались на следующем этапе для обучения U-образной нейросети для шумоподавления. Генеративный компонент U-образной нейросети для шумоподавления применял функцию потерь для предсказания шума, характерную для моделей семейства Stable Diffusion. Размерность латентного пространства была выбрана на основе предварительных экспериментов с архитектурой Stable Diffusion 3.5 Large.

Оценка эффективности метода. Эффективность разработанного метода оценивалась с использованием нескольких количественных метрик, направленных на измерение степени согласованности между синтезированной речью, мимикой и жестами, с учетом эмоциональной окраски текста. Особое внимание уделялось метрикам оценки способности нейросетевой модели передавать эмоции через визуальные проявления, что является ключевым фактором в создании правдоподобных цифровых аватаров. Кроме того, оценивался уровень артефактов, возникающих при генерации — таких как неестественные движения, десинхронизация аудио- и видеосигналов, а также ошибки в лицевой анимации. Наконец оценивались показатели синхронизации и качества визуального ряда, а также интегральные оценки аудиовизуальной согласованности. В таблице приведены результаты оценивания BiMoDiCA и абляционный анализ эффективности компонентов предложенного метода.

Результаты показывают (таблица), что исключение **L** (векторное представление ключевых точек) приводит к умеренному ухудшению визуальных характеристик. Значения метрик FID и FVD выросли с 50,13 до 56,24 и с 601,70 до 630,39 соответственно, а показатели метрик синхронизации артикуляции и согласованности движений несколько снизились. При этом наиболее выраженное ухудшение качества наблюдается при исключении $\mathbf{z_A}$ (векторное представление аудиосигнала), что сопровождается наибольшим ростом FID до 66,83 и FVD до 737,51, а также снижением интегральных метрик синхронизации (Sync-C до 5,141 и CSIM до 0,427). Это подчеркивает ключевую роль акустического энкодера в обеспечении достоверной синхронизации аудиоречи с визуальными проявлениями и мимикой, а также в восприятии эмоциональной окраски. Исключение $\mathbf{z_E}$ (векторное представление эмоций) оказало существенное влияние на способность модели воспроизводить эмоциональные состояния. Рост E-FID с 2,226 до 3,183 и снижение Sync-C с 6,398 до 5,628 свидетельствуют об уменьшении согласованности между эмоциональной окраской текста, синтезированной речью и визуальными проявлениями. Удаление $\mathbf{z_T}$ (векторное представление текста) в большей степени затронуло показатели динамической целостности визуального ряда и синхронизации движений, что проявилось в росте FVD до 685,27 и снижении Sync-C до 5,921, однако влияние на эмоциональные показатели оказалось менее выраженным по сравнению с $\mathbf{z_E}$. Таким образом, результаты

Таблица. Результаты оценивания метода BiMoDiCA
Table. Evaluation results of the BiMoDiCA method

Метрики	BiMoDiCA	Без **L**	Без $\mathbf{z_E}$	Без $\mathbf{z_T}$	Без $\mathbf{z_A}$
FID↓ [36]	50,130	56,240	60,350	58,210	66,830
FVD↓ [36]	601,700	630,390	715,880	685,270	737,510
SSIM↑ [37]	0,752	0,735	0,711	0,723	0,684
PSNR↑ [38]	21,997	21,512	20,678	21,120	19,839
E-FID↓ [39]	2,226	2,461	3,183	2,807	3,556
Sync-D↓ [40]	7,003	7,425	8,486	7,844	9,027
Sync-C↑ [40]	6,398	6,182	5,628	5,921	5,141
HKC↑ [17]	0,912	0,895	0,859	0,880	0,852
HKV↑ [17]	25,020	23,790	21,840	22,570	19,980
CSIM↑ [17]	0,518	0,501	0,466	0,483	0,427

Примечание. ↑ и ↓ — лучшие результаты соответствуют большим и меньшим значениям.

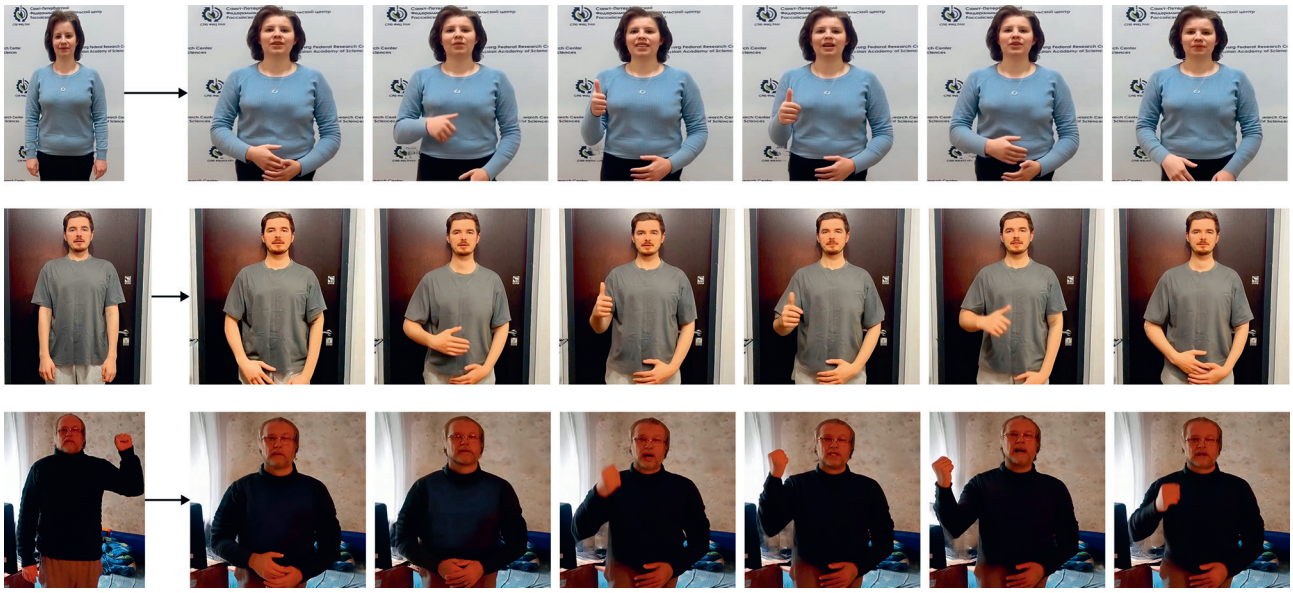

Рис. 2. Примеры ключевых кадров генерации цифровых аватаров

Fig. 2. The following examples are provided to illustrate keyframes in the context of digital avatar generation

абляционного анализа демонстрируют взаимосвязанную структуру компонентов метода и подтверждают, что исключение одного из компонентов приводит к ухудшению эффективности предложенного метода.

На рис. 2 представлены ключевые кадры, иллюстрирующие визуальные результаты генерации цифровых аватаров на основе эмоционального текста и синтезированной речи. Демонстрируется высокая согласованность мимики, жестикуляции и аудиоречи, подтверждающая эффективность предложенного метода. Дополнительные примеры генерации видеоконтента доступны на персональной странице метода[1].

Важно отметить, что в рамках данного исследования акцент был сделан на эмоциональных жестах, характерных для русского жестового языка, а также на интеграции управляющих жестов. Сравнение с аналогичными методами не проводилось, поскольку в существующих научных публикациях нет аналогичных решений, учитывающих синхронизацию эмоций, жестов и речи с учетом особенностей русского жестового языка и управления жестами для умных систем.

Заключение

В данной работе представлен метод генерации анимации цифрового аватара под названием BiMoDiCA с синхронной речевой и невербальной экспрессией на основе бимодального ввода — изображения и эмоционально окрашенного текста. BiMoDiCA основан на диффузионной генеративной архитектуре, интегрирующей акустические, визуальные и аффективные признаки. Такая интеграция обеспечивает реалистичную синхронизацию между речью, артикуляцией, мимикой и жестикуляцией, а также адаптацию поведения аватара в соответствии с эмоциональным контекстом.

Проведенные экспериментальные исследования подтвердили высокую эффективность предлагаемого метода, который показал высокие значения по ряду ключевых метрик: FID (50,13) и FVD (601,70) подтвердили реалистичность видеоконтента и его динамическую устойчивость; SSIM (0,752) и PSNR (21,997) отразили высокое структурное сходство и качество изображения; E-FID (2,226) зафиксировала достоверную передачу эмоционального содержания; метрики синхронизации Sync-D (7,003) и Sync-C (6,398) продемонстрировали согласованность аудиовизуальных данных, в то время как CSIM (0,518) и значения HKC/HKV (0,912/25,02) подтвердили точность поз и вариативность жестов. Эти результаты свидетельствуют о способности метода создавать выразительные и эмоционально насыщенные анимации с высокой степенью когерентности между модальностями.

Особый вклад работы заключается в учете особенностей русского жестового языка при формировании невербальных шаблонов, что расширяет сферу применения метода в инклюзивных системах взаимодействия и аффективно-ориентированных интерфейсах.

Будущие направления исследований включают развитие компонента синтеза речи за счет генерации эмоционально окрашенной и просодически выразительной аудиоречи, устранение артефактов, возникающих на элементах одежды и текстурах при генерации видеоконтента, а также оптимизацию нейросетевых моделей для эффективного внедрения в устройства с ограниченными вычислительными ресурсами. Кроме того, предполагается изучение применения предлагаемого метода в задачах длительного аффективного мониторинга и персонализированного взаимодействия, ориентированного на создание адаптивных, эмоционально чувствительных и эмпатичных цифровых агентов.

[1] BiMoDiCA [Электронный ресурс]. Режим доступа: https://smil-spcras.github.io/BiMoDiCA, свободный. Яз. рус. (дата обращения: 21.04.2025).

Литература

1. Sincan O.M., Keles H.Y. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods // IEEE Access. 2020. V. 8. P. 181340–181355. https://doi.org/10.1109/ACCESS.2020.3028072
2. Kapitanov A., Kvanchiani K., Nagaev A., Kraynov R., Makhliarchuk A. HaGRID-HAnd Gesture Recognition Image Dataset // Proc. of the Winter Conference on Applications of Computer Vision (WACV). 2024. P. 4560–4569. https://doi.org/10.1109/WACV57701.2024.00451
3. Busso C., Bulut M., Lee C.C., Kazemzadeh A., Mower E., Kim S., Chang J., Lee S., Narayanan S.S. IEMOCAP: interactive emotional dyadic motion capture database // Language Resources and Evaluation. 2008. V. 42. N 4. P. 335–359. https://doi.org/10.1007/s10579-008-9076-6
4. Shen K., Guo C., Kaufmann M., Zarate J., Valentin J., Song J., Hilliges O. X-Avatar: expressive human avatars // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 16911–16921. https://doi.org/10.1109/CVPR52729.2023.01622
5. Zhang H., Chen B., Yang H., Qu L., Wang X., Chen L., Long C., Zhu F., Du D., Zheng M. AvatarVerse: high-quality and stable 3D avatar creation from text and pose // Proc. of the AAAI Conference on Artificial Intelligence. 2024. V. 38. N 7. P. 7124–7132. https://doi.org/10.1609/aaai.v38i7.28540
6. Kim K., Song B. Robust 3D human avatar reconstruction from monocular videos using depth optimization and camera pose estimation // IEEE Access. 2025. V. 13. P. 57886–57897. https://doi.org/10.1109/ACCESS.2025.3556445
7. Yuan Y., Li X., Huang Y., De Mello S., Nagano K., Kautz J., Iqbal U. Gavatar: animatable 3D gaussian avatars with implicit mesh learning // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 896–905. https://doi.org/10.1109/CVPR52733.2024.00091
8. Teotia K., Mallikarjun B.R., Pan X., Kim H., Garrido P., Elgharib M., Theobalt C. HQ3DAvatar: high-quality implicit 3D head avatar // ACM Transactions on Graphics. 2024. V. 43. N 3. P 1–24. https://doi.org/10.1145/3649889
9. Yang L., Zhang Z., Song Y., Hong S., Xu R., Zhao Y., Zhang W., Cui B., Yang M. Diffusion models: a comprehensive survey of methods and applications // ACM Computing Surveys. 2023. V. 56. N 4. P. 1–39. https://doi.org/10.1145/3626235
10. Karras J., Holynski A., Wang T., Kemelmacher-Shlizerman I. DreamPose: fashion image-to-video synthesis via stable diffusion // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023. P. 22623–22633. https://doi.org/10.1109/ICCV51070.2023.02073
11. Huang Z., Tang F., Zhang Y., Cun X., Cao J., Li J., Lee T. Make-Your-Anchor: a diffusion-based 2D avatar generation framework // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 6997–7006. https://doi.org/10.1109/CVPR52733.2024.00668
12. Blattmann A., Dockhorn T., Kulal S., Mendelevitch D., Kilian M., Lorenz D., Levi Y., English Z., Voleti V., Letts A., Jampani V., Rombach R. Stable video diffusion: scaling latent video diffusion models to large datasets // arXiv. 2023. arXiv:2311.15127. https://doi.org/10.48550/arXiv.2311.15127
13. Zhang L., Rao A., Agrawala M. Adding conditional control to text-to-image diffusion models // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023. P. 3813–3824. https://doi.org/10.1109/ICCV51070.2023.00355
14. Zhuang S., Li K., Chen X., Wang Y., Liu Z., Qiao Y., Wang Y. Vlogger: make your dream a vlog // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 8806–8817. https://doi.org/10.1109/CVPR52733.2024.00841
15. Xu M., Li H., Su Q., Shang H., Zhang L., Liu C., Wang J., Yao Y., Zhu S. Hallo: hierarchical audio-driven visual synthesis for portrait image animation // arXiv. 2024. arXiv:2406.08801. https://doi.org/10.48550/arXiv.2406.08801
16. Yang S., Li H., Wu J., Jing M., Li L., Ji R., Liang J., Fan H., Wang J. MegActor-Sigma: unlocking flexible mixed-modal control in portrait animation with diffusion transformer // Proc. of the AAAI Conference on Artificial Intelligence. 2025. V.39. N 9. P. 9256–9264. https://doi.org/10.1609/aaai.v39i9.33002

References

1. Sincan O.M., Keles H.Y. AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods. *IEEE Access*, 2020, vol. 8, pp. 181340–181355. https://doi.org/10.1109/ACCESS.2020.3028072
2. Kapitanov A., Kvanchiani K., Nagaev A., Kraynov R., Makhliarchuk A. HaGRID-HAnd Gesture Recognition Image Dataset. *Proc. of the Winter Conference on Applications of Computer Vision (WACV)*. 2024, pp. 4560–4569. https://doi.org/10.1109/WACV57701.2024.00451
3. Busso C., Bulut M., Lee C.C., Kazemzadeh A., Mower E., Kim S., Chang J., Lee S., Narayanan S.S. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 2008, vol. 42, no. 4, pp. 335–359. https://doi.org/10.1007/s10579-008-9076-6
4. Shen K., Guo C., Kaufmann M., Zarate J., Valentin J., Song J., Hilliges O. X-Avatar: expressive human avatars. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16911–16921. https://doi.org/10.1109/CVPR52729.2023.01622
5. Zhang H., Chen B., Yang H., Qu L., Wang X., Chen L., Long C., Zhu F., Du D., Zheng M. AvatarVerse: high-quality and stable 3D avatar creation from text and pose. *Proc. of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 7, pp. 7124–7132. https://doi.org/10.1609/aaai.v38i7.28540
6. Kim K., Song B. Robust 3D human avatar reconstruction from monocular videos using depth optimization and camera pose estimation. *IEEE Access*, 2025, vol. 13, pp. 57886–57897. https://doi.org/10.1109/ACCESS.2025.3556445
7. Yuan Y., Li X., Huang Y., De Mello S., Nagano K., Kautz J., Iqbal U. Gavatar: animatable 3D gaussian avatars with implicit mesh learning. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 896–905. https://doi.org/10.1109/CVPR52733.2024.00091
8. Teotia K., Mallikarjun B.R., Pan X., Kim H., Garrido P., Elgharib M., Theobalt C. HQ3DAvatar: high-quality implicit 3D head avatar. *ACM Transactions on Graphics*, 2024, vol. 43, no. 3, pp 1–24. https://doi.org/10.1145/3649889
9. Yang L., Zhang Z., Song Y., Hong S., Xu R., Zhao Y., Zhang W., Cui B., Yang M. Diffusion models: a comprehensive survey of methods and applications. *ACM Computing Surveys*, 2023, vol. 56, no. 4, pp. 1–39. https://doi.org/10.1145/3626235
10. Karras J., Holynski A., Wang T., Kemelmacher-Shlizerman I. DreamPose: fashion image-to-video synthesis via stable diffusion. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22623–22633. https://doi.org/10.1109/ICCV51070.2023.02073
11. Huang Z., Tang F., Zhang Y., Cun X., Cao J., Li J., Lee T. Make-Your-Anchor: a diffusion-based 2D avatar generation framework. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 6997–7006. https://doi.org/10.1109/CVPR52733.2024.00668
12. Blattmann A., Dockhorn T., Kulal S., Mendelevitch D., Kilian M., Lorenz D., Levi Y., English Z., Voleti V., Letts A., Jampani V., Rombach R. Stable video diffusion: scaling latent video diffusion models to large datasets. *arXiv*, 2023, arXiv:2311.15127. https://doi.org/10.48550/arXiv.2311.15127
13. Zhang L., Rao A., Agrawala M. Adding conditional control to text-to-image diffusion models. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3813–3824. https://doi.org/10.1109/ICCV51070.2023.00355
14. Zhuang S., Li K., Chen X., Wang Y., Liu Z., Qiao Y., Wang Y. Vlogger: make your dream a vlog. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 8806–8817. https://doi.org/10.1109/CVPR52733.2024.00841
15. Xu M., Li H., Su Q., Shang H., Zhang L., Liu C., Wang J., Yao Y., Zhu S. Hallo: hierarchical audio-driven visual synthesis for portrait image animation. *arXiv*, 2024, arXiv:2406.08801. https://doi.org/10.48550/arXiv.2406.08801
16. Yang S., Li H., Wu J., Jing M., Li L., Ji R., Liang J., Fan H., Wang J. MegActor-Sigma: unlocking flexible mixed-modal control in portrait animation with diffusion transformer. *Proc. of the AAAI Conference on Artificial Intelligence*, 2025, vol.39, no. 9, pp. 9256–9264. https://doi.org/10.1609/aaai.v39i9.33002

17. Lin G., Jiang J., Liang C., Zhong T., Yang J., Zheng Y. CyberHost: taming audio-driven avatar diffusion model with region codebook attention // arXiv. 2024. arXiv:2409.01876. https://doi.org/10.48550/arXiv.2409.01876
18. Chen Z., Cao J., Chen Z., Li Y., Ma C. EchoMimic: lifelike audio-driven portrait animations through editable landmark conditions // Proc. of the AAAI Conference on Artificial Intelligence. 2025. V. 39. N 3. P. 2403–2410. https://doi.org/10.1609/aaai.v39i3.32241
19. Serengil S., Özpınar A. A benchmark of facial recognition pipelines and co-usability performances of modules // Bilişim Teknolojileri Dergisi. 2024. V. 17. N 2. P. 95–107. https://doi.org/10.17671/gazibtd.1399077
20. Bazarevsky V., Kartynnik Y., Vakunov A., Raveendran K., Grundmann M. Blazeface: sub-millisecond seural face detection on mobile GPUs // arXiv. 2019. arXiv:1907.05047. https://doi.org/10.48550/arXiv.1907.05047
21. Zhang F., Bazarevsky V., Vakunov A., Tkachenka A., Sung G., Chang C.L., Grundmann M. MediaPipe hands: on-device real-time hand tracking // arXiv. 2020. arXiv:2006.10214. https://doi.org/10.48550/arXiv.2006.10214
22. Bazarevsky V., Grishchenko I., Raveendran K., Zhu T., Zhang F., Grundmann M. BlazePose: on-device real-time body pose tracking // arXiv. 2020. arXiv:2006.10204. https://doi.org/10.48550/arXiv.2006.10204
23. Xu J., Zou X., Huang K., Chen Y., Liu B., Cheng M., Shi X., Huang J. EasyAnimate: a high-performance long video generation method based on transformer Architecture // arXiv. 2024. arXiv:2405.18991. https://doi.org/10.48550/arXiv.2405.18991
24. Hu L. Animate anyone: consistent and controllable image-to-video synthesis for character animation // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 8153–8163. https://doi.org/10.1109/CVPR52733.2024.00779
25. Ryumina E., Ryumin D., Axyonov A., Ivanko D., Karpov A. Multi-corpus emotion recognition method based on cross-modal gated attention fusion // Pattern Recognition Letters. 2025. V. 190. P. 192–200. https://doi.org/10.1016/j.patrec.2025.02.024
26. Peng Y., Sudo Y., Shakeel M., Watanabe S. OWSM-CTC: an open encoder-only speech foundation model for speech recognition, translation, and language identification // Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024. V. 1. P. 10192–10209. https://doi.org/10.18653/v1/2024.acl-long.549
27. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I. Attention is all you need // Proc. of the Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. P. 1–11.
28. Kapitanov A., Kvanchiani K., Nagaev A., Petrova E. Slovo: Russian sign language dataset // Lecture Notes in Computer Science. 2023. V. 14253. P. 63–73. https://doi.org/10.1007/978-3-031-44137-0_6
29. Xie L., Wang X., Zhang H., Dong C., Shan Y. VFHQ: a high-quality dataset and benchmark for video face super-resolution // IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. P. 657–665. https://doi.org/10.1109/CVPRW56347.2022.00081
30. Kagirov I., Ivanko D., Ryumin D., Axyonov A., Karpov A. TheRuSLan: database of russian sign language // Proc. of the12th Conference on Language Resources and Evaluatio (LREC). 2020. P. 6079–6085.
31. Кагиров И.А., Рюмин Д.А., Аксенов А.А., Карпов А.А. Мультимедийная база данных жестов русского жестового языка в трехмерном формате // Вопросы языкознания. 2020. № 1.С. 104-123. https://doi.org/10.31857/S0373658X0008302-1
32. Axyonov A., Ryumin D., Ivanko D., Kashevnik A., Karpov A. Audio-visual speech recognition in-the-wild: multi-angle vehicle cabin corpus and attention-based method // Proc. of the 49th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2024. P. 8195–8199. https://doi.org/10.1109/ICASSP48485.2024.10448048
33. Liu Z. Super Convergence cosine annealing with warm-up learning rate // Proc. of the 2nd International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA). 2022. P. 1–7.
34. Wang P., Shen L., Tao Z., He S., Tao D. Generalization analysis of stochastic weight averaging with general sampling // Proc. of the 41st International Conference on Machine Learning (ICML). 2024. P. 51442–51464.
35. Yang H., Zhang Z., Tang H., Qian J., Yang J. ConsistentAvatar: learning to diffuse fully consistent talking head avatar with temporal guidance // Proc. of the 32nd ACM International Conference on

17. Lin G., Jiang J., Liang C., Zhong T., Yang J., Zheng Y. CyberHost: taming audio-driven avatar diffusion model with region codebook attention. *arXiv*, 2024, arXiv:2409.01876. https://doi.org/10.48550/arXiv.2409.01876
18. Chen Z., Cao J., Chen Z., Li Y., Ma C. EchoMimic: lifelike audio-driven portrait animations through editable landmark conditions. *Proc. of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, no. 3, pp. 2403–2410. https://doi.org/10.1609/aaai.v39i3.32241
19. Serengil S., Özpınar A. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilişim Teknolojileri Dergisi*, 2024, vol. 17, no. 2, pp. 95–107. https://doi.org/10.17671/gazibtd.1399077
20. Bazarevsky V., Kartynnik Y., Vakunov A., Raveendran K., Grundmann M. Blazeface: sub-millisecond seural face detection on mobile GPUs. *arXiv*, 2019, arXiv:1907.05047. https://doi.org/10.48550/arXiv.1907.05047
21. Zhang F., Bazarevsky V., Vakunov A., Tkachenka A., Sung G., Chang C.L., Grundmann M. MediaPipe hands: on-device real-time hand tracking. *arXiv*, 2020, arXiv:2006.10214. https://doi.org/10.48550/arXiv.2006.10214
22. Bazarevsky V., Grishchenko I., Raveendran K., Zhu T., Zhang F., Grundmann M. BlazePose: on-device real-time body pose tracking. *arXiv*, 2020, arXiv:2006.10204. https://doi.org/10.48550/arXiv.2006.10204
23. Xu J., Zou X., Huang K., Chen Y., Liu B., Cheng M., Shi X., Huang J. EasyAnimate: a high-performance long video generation method based on transformer Architecture. *arXiv*, 2024, arXiv:2405.18991. https://doi.org/10.48550/arXiv.2405.18991
24. Hu L. Animate anyone: consistent and controllable image-to-video synthesis for character animation. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 8153–8163. https://doi.org/10.1109/CVPR52733.2024.00779
25. Ryumina E., Ryumin D., Axyonov A., Ivanko D., Karpov A. Multi-corpus emotion recognition method based on cross-modal gated attention fusion. *Pattern Recognition Letters*, 2025, vol. 190, pp. 192–200. https://doi.org/10.1016/j.patrec.2025.02.024
26. Peng Y., Sudo Y., Shakeel M., Watanabe S. OWSM-CTC: an open encoder-only speech foundation model for speech recognition, translation, and language identification. *Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, vol. 1, pp. 10192–10209. https://doi.org/10.18653/v1/2024.acl-long.549
27. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., Polosukhin I. Attention is all you need. *Proc. of the Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 1–11.
28. Kapitanov A., Kvanchiani K., Nagaev A., Petrova E. Slovo: Russian sign language dataset. *Lecture Notes in Computer Science*, 2023, vol. 14253, pp. 63–73. https://doi.org/10.1007/978-3-031-44137-0_6
29. Xie L., Wang X., Zhang H., Dong C., Shan Y. VFHQ: a high-quality dataset and benchmark for video face super-resolution. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 657–665. https://doi.org/10.1109/CVPRW56347.2022.00081
30. Kagirov I., Ivanko D., Ryumin D., Axyonov A., Karpov A. TheRuSLan: database of russian sign language. *Proc. of the12th Conference on Language Resources and Evaluatio (LREC)*, 2020. pp. 6079–6085.
31. Kagirov I., Ryumin D.A., Axyonov A.A., Karpov A.A. Multimedia database of russian sign language items in 3D. *Voprosy jazykoznanija*, 2020, no. 1., pp. 104–123. (in Russian). https://doi.org/10.31857/S0373658X0008302-1
32. Axyonov A., Ryumin D., Ivanko D., Kashevnik A., Karpov A. Audio-visual speech recognition in-the-wild: multi-angle vehicle cabin corpus and attention-based method. *Proc. of the 49th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024, pp. 8195–8199. https://doi.org/10.1109/ICASSP48485.2024.10448048
33. Liu Z. Super Convergence cosine annealing with warm-up learning rate. *Proc. of the 2nd International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, 2022, pp. 1–7.
34. Wang P., Shen L., Tao Z., He S., Tao D. Generalization analysis of stochastic weight averaging with general sampling. *Proc. of the 41st International Conference on Machine Learning (ICML)*, 2024, pp. 51442–51464.
35. Yang H., Zhang Z., Tang H., Qian J., Yang J. ConsistentAvatar: learning to diffuse fully consistent talking head avatar with temporal

Multimedia. 2024. P. 3964–3973. https://doi.org/10.1145/3664647.3680619
36. Unterthiner T., Van Steenkiste S., Kurach K., Marinier R., Michalski M., Gelly S. Towards accurate generative models of video: a new metric and challenges // arXiv. 2018. arXiv:1812.01717. https://doi.org/10.48550/arXiv.1812.01717
37. Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity // IEEE Transactions on Image Processing. 2004. V. 13. N 4. P. 600–612. https://doi.org/10.1109/TIP.2003.819861
38. Hore A., Ziou D. Image quality metrics: PSNR vs. SSIM // Proc. of the 20th International Conference on Pattern Recognition. 2010. P. 2366–2369. https://doi.org/10.1109/ICPR.2010.579
39. Deng Y., Yang J., Xu S., Chen D., Jia Y., Tong X. Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set // Proc. of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 285–295. https://doi.org/10.1109/CVPRW.2019.00038
40. Prajwal K.R., Mukhopadhyay R., Namboodiri V.P., Jawahar C.V. A lip sync expert is all you need for speech to lip generation in the wild // Proc. of the 28th ACM International Conference on Multimedia. 2020. P. 484–492. https://doi.org/10.1145/3394171.3413532

guidance. *Proc. of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3964–3973. https://doi.org/10.1145/3664647.3680619
36. Unterthiner T., Van Steenkiste S., Kurach K., Marinier R., Michalski M., Gelly S. Towards accurate generative models of video: a new metric and challenges. *arXiv*, 2018, arXiv:1812.01717. https://doi.org/10.48550/arXiv.1812.01717
37. Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004, vol. 13, no. 4, pp. 600–612. https://doi.org/10.1109/TIP.2003.819861
38. Hore A., Ziou D. Image quality metrics: PSNR vs. SSIM. *Proc. of the 20th International Conference on Pattern Recognition*, 2010, pp. 2366–2369. https://doi.org/10.1109/ICPR.2010.579
39. Deng Y., Yang J., Xu S., Chen D., Jia Y., Tong X. Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set. *Proc. of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 285–295. https://doi.org/10.1109/CVPRW.2019.00038
40. Prajwal K.R., Mukhopadhyay R., Namboodiri V.P., Jawahar C.V. A lip sync expert is all you need for speech to lip generation in the wild. *Proc. of the 28th ACM International Conference on Multimedia*, 2020, pp. 484–492. https://doi.org/10.1145/3394171.3413532

Авторы

Аксёнов Александр Александрович — кандидат технических наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57203963345, https://orcid.org/0000-0002-7479-2851, axyonov.a@iias.spb.su

Рюмина Елена Витальевна — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57220572427, https://orcid.org/0000-0002-4135-6949, ryumina.e@iias.spb.su

Рюмин Дмитрий Александрович — кандидат технических наук, старший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, SC 57191960214, https://orcid.org/0000-0002-7935-0569, ryumin.d@iias.spb.su

Authors

Alexander A. Axyonov — PhD, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 57203963345, https://orcid.org/0000-0002-7479-2851, axyonov.a@iias.spb.su

Elena V. Ryumina — Junior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 57220572427, https://orcid.org/0000-0002-4135-6949, ryumina.e@iias.spb.su

Dmitry A. Ryumin — PhD, Senior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, SC 57191960214, https://orcid.org/0000-0002-7935-0569, ryumin.d@iias.spb.su

Статья поступила в редакцию 24.04.2025
Одобрена после рецензирования 10.06.2025
Принята к печати 25.07.2025

Received 24.04.2025
Approved after reviewing 10.06.2025
Accepted 25.07.2025