

Captain Safari: Мировой движок

Yu-ChengChou¹ Xingrui Wang¹ Yitong Li² Jiahao Wang¹ Hanting Liu¹CihangXie³ Alan Yuille¹ Junfei Xiao^{1✉}
 Университет Джонса Хопкинса² Цинхуа университет³ Университет Калифорний в
 Санта-Крузе <https://johnson111788.github.io/open-safari/>

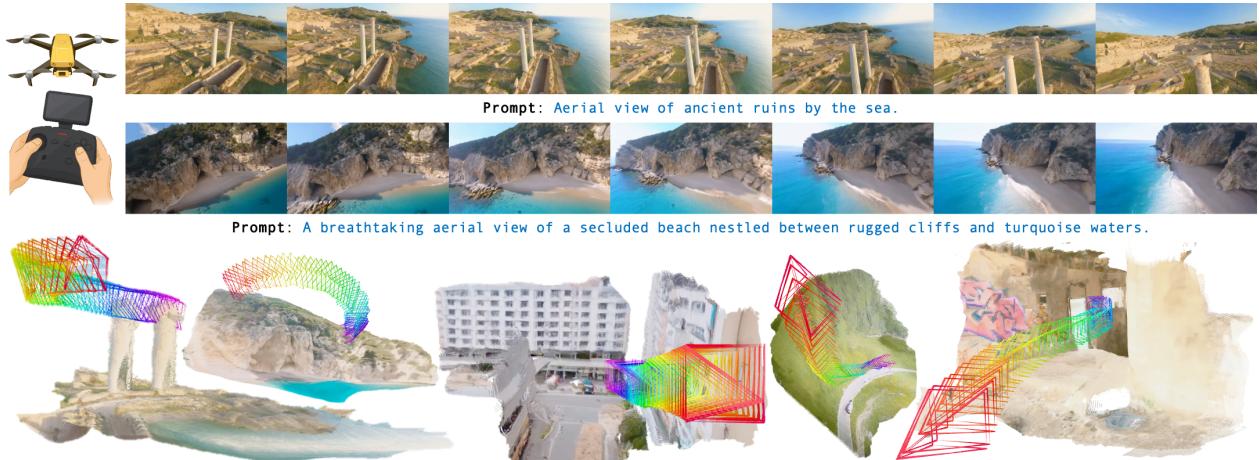


Рисунок 1. CaptainSafari — это мировой движок, учитывающий позу, который генерирует FPV видео с длинным горизонтом и 3D-согласованностью из любой заданной пользователем траектории камеры. Извлекая память мира, согласованную с позой, он сохраняет стабильность геометрии при значительных изменениях точки обзора и восстанавливает четкие, хорошо сформированные структуры, точно отслеживая агрессивное движение 6-DoF.

Аннотация

Мировые движки нацелены на синтез длинных, 3D-согласованных видео, поддерживающих интерактивное исследование сцен при управляемом пользователем движении камеры. Однако существующие системы испытывают трудности при агрессивных траекториях 6-DoF и сложных уличных планировках: они теряют долгосрочную геометрическую согласованность, отклоняются от целевого пути или сворачиваются в чрезмерно консервативное движение. Для этого мы представляем Captain Safari, мировой движок, обусловленный позой, который генерирует видео, извлекая из постоянной памяти мира. Учитывая путь камеры, наш метод поддерживает динамическую локальную память и использует извлекатель для получения токенов мира, согласованных с позой, которые затем обуславливают генерацию видео вдоль траектории. Этот дизайн позволяет модели поддерживать стабильную 3D-структуру, точно выполняя сложные маневры камеры.

Для оценки этой настройки мы создаем OpenSafari, новый FPV набор данных в естественных условиях, содержащий высокодинамичные видео с дронов с проверенными траекториями камеры, построенный через многоглавую геометрическую и кинематическую валидацию.

В отношении качества видео, 3D согласованности и следования траектории CaptainSafari значительно превосходит современные генераторы с управлением камерой. Он снижает MEt3R с 0.3703 до 0.3690, улучшает AUC@30 с 0.181 до 0.200 и дает значительно более низкий FVD, чем все базовые модели с управлением камерой. Более важно, в исследовании с участием 50 человек, где аннотаторы выбирали лучший результат среди пяти анонимных моделей, 67,6% предпочтений отдают нашему методу по всем осмым. Наши результаты демонстрируют, что поза-обусловленная память мира является мощным механизмом для долгосрочной, управляемой генерации видео и предоставляют OpenSafari как сложный новый бенчмарк для будущих исследований мировых движков.

1. Введение

Симуляция согласованных 3D миров через управляемую генерацию видео долгое время была основным вызовом для дополненной реальности, воплощенного ИИ и виртуальных агентов [8–10, 13–16, 19, 20, 23, 26, 29, 40, 43–45, 50, 51, 57]. Классические игровые движки и физические симуляторы предлагают явную геометрию и точный контроль, но требуют значительных ручных усилий и дорогих вычислений [7, 28, 32]. Более-

[✉] Автор для переписки: Junfei Xiao (xiaojf97@gmail.com)
 Предварительная версия, работа в процессе.

того, хотя они могут достигать визуального реализма в специализированных областях, они все же не способны передать богатство и разнообразие, характерные для реального мира, такие как природные сцены [27, 35, 58]. В отличие от них, современные модели диффузии видео синтезируют высококачественные, разнообразные видео из текста или изображений, но обычно работают как генераторы клипов без постоянного состояния мира: *они испытывают трудности с долгосрочной 3D согласованностью, сложным следованием траектории и точной реконструкцией разнообразных сцен* [18, 24]. В этой работе мы стремимся преодолеть этот разрыв с помощью *Captain Safari*, мирового движка, который позволяет моделировать 3D-согласованные и разнообразные среды, превосходя ограничения традиционных игровых движков в плане универсальности, разнообразия и интерактивности.

Современные модели видеомиров сталкиваются с тремя взаимосвязанными проблемами. Во-первых, *долгосрочная согласованность* ограничена временными окнами контекстных кадров; модели часто «забывают» удаленные пейзажи или нарушают пространственную согласованность, что приводит к резким изменениям внешнего вида, нарушающим реализм и непрерывность создаваемой среды [8, 14, 45]. Во-вторых, достижение *сложных маневров камеры при строгой 3D согласованности* остается сложной задачей: существующие методы, основанные на позе или траектории, обычно хорошо работают только для медленных, почти прямолинейных движений [16, 34, 49]. Когда путь включает быстрое движение 6-DoF, сильный параллакс или резкие повороты, модели демонстрируют компромисс — либо ослабляют движение и ограничивают изменения точки зрения для сохранения геометрии, либо следят запрашиваемому пути ценой искажений, мерцания и структурного дрейфа. В-третьих, текущие подходы недостаточно представляют *сложные уличные макеты*. Большая часть работ сосредоточена на структурированных, ограниченных условиях (например, внутренние туры, сцены вождения или видео о недвижимости), и модели редко подвергаются стресс-тестированию в условиях FPV на открытом воздухе, где камера маневрирует вокруг зданий, растительности и разнообразного рельефа с существенным параллаксом [6, 22, 59, 60]. В результате методы, которые выглядят конкурентоспособными в упрощенных условиях, часто не могут сохранить геометрию и внешний вид при столкновении с действительно разнообразными, сложными уличными сценами.

Чтобы решить эти проблемы, мы представляем *Captain Safari*, поза-осведомленный мировой движок, который явно поддерживает постоянное представление состояния мира для обеспечения *долгосрочной 3D согласованности* при сильном параллаксе. Поскольку хранение и распространение полного долгосрочного состояния вычислительно затратно, мы разработали механизм извлечения, который *выбирает и агрегирует* только наиболее информативные подсказки сцены, тем самым обеспечивая сильное геометрическое руководство без чрезмерных затрат. Важно, что это извлечение *поза-осведомленное*: учитывая целевую позу камеры, оно собирает выровненный по позе мировой приоритет, который направляет процесс генерации, позволяя точно отслеживать *агрессивные маневры камеры*, сохраняя при этом 3D-согласованную структуру в сложных средах.

Кроме того, чтобы устранить разрыв в *сложных уличных макетах* и *агрессивном движении камеры*, мы создаем *OpenSafari*,

крупномасштабный набор данных высокодинамичных FPV видео с дронов с проверенными позами камеры. Большая часть литературы ориентирована на структурированные, ограниченные условия (например, внутренние туры, вождение или видео о недвижимости), и даже уличные наборы данных обычно содержат медленное, почти прямолинейное движение. В отличие от них, *OpenSafari* включает FPV полеты в дикой природе, которые маневрируют вокруг зданий и растительности на неровной местности, демонстрируя большой параллакс, быстрые маневры 6-DoF и резкие изменения точки зрения. В сочетании с проверенными траекториями камеры, эти видео представляют разнообразные, загроможденные уличные сцены и дальние движения, бросая вызов моделям в поддержании 3D согласованности при точном отслеживании сложных маневров.

Мы оцениваем *Captain Safari* по трем осиам: *качество видео, 3D согласованность и следование траектории*. По всем этим критериям наш метод стабильно превосходит современные генераторы видео с управлением камерой на *OpenSafari*: Таблица 1 показывает явные улучшения в 3D согласованности и точном отслеживании при сложных маневрах, сохранив при этом высокое перцептивное качество. Важно, что крупномасштабное исследование с участием людей (Таблица 2) показывает, что *Captain Safari* получает 67% голосов в пятисторонних сравнениях, что указывает на то, что улучшения заметны на перцептивном уровне. Качественные сравнения (Рис. 4 и Рис. 5) дополнительно демонстрируют стабильную геометрию при дальних маршрутах и точное следование резким поворотам камеры с 6-DoF в загроможденных уличных сценах.

В заключение, наши вклады следующие:

1. Мы представляем *Captain Safari*, первый метод генерации вideo с управлением камерой, обеспечивающий долгосрочную 3D согласованность при отслеживании агрессивных маневров FPV.
2. Мы предлагаем *поисковую систему, управляемую позой*, для долгосрочного извлечения, которая эффективно сочетает строгую 3D согласованность с точным отслеживанием сложных маневров.
3. Мы создаем *OpenSafari*, крупномасштабный FPV набор данных в естественных условиях с проверенными позами камеры, включающий разнообразные, загроможденные уличные сцены и быстрое движение 6-DoF, что проверяет управление камерой на согласованность геометрии.
4. В *OpenSafari* наше извлечение, учитывающее позу, значительно улучшает качество видео, 3D согласованность и выравнивание траектории, также достигая 67% уровня предпочтения среди людей.

2. Связанные работы

2.1. 3D-согласованные мировые модели

Ранние подходы к преобразованию изображений в 3D восстанавливают геометрию косвенно через многовидовую согласованность или неявные поля, но часто не могут поддерживать целостную структуру при значительных изменениях вида [13, 21, 43, 50]. Недавние усилия интегрируют 3D-рассуждения в процесс генерации. DiffusionGS [5] внедряет Gaussian Splatting в диффузионный денойзер, обеспечивая согласованность видов и позволяя одностадийную, масштабируемую 3D-генерацию. GenEx [23] и EvoWorld [40] расширяют эту идею от статической реконструкции до создания динамичных миров, генерируя исследуемые панорамные 360° окружения.

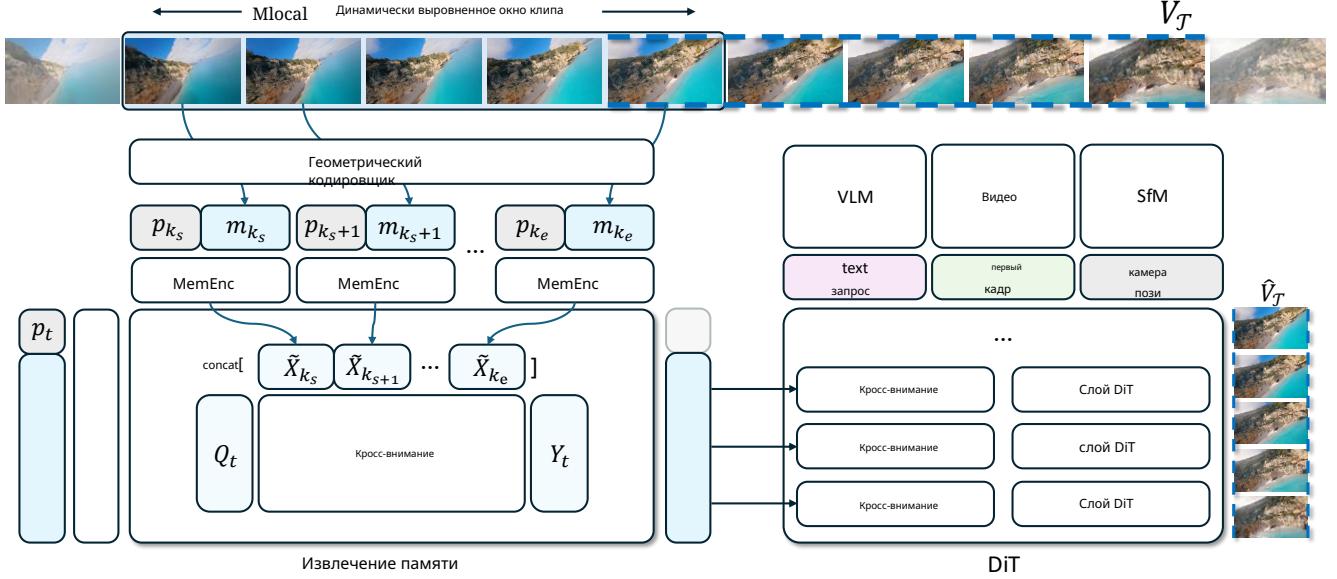


Рисунок 2. Обзор метода. *Captain Safari* создает локальную память мира и, имея заданную позу камеры, извлекает выровненные по позе токены, которые резюмируют сцену. Эти токены затем определяют генерацию видео вдоль заданной пользователем траектории, сохраняя стабильную 3D компоновку.

основанные на физических предпосылках. В дополнение к этим генеративным реконструкциям, Geometry Forcing [44] и Memory Forcing [14] явно связывают обучающие сигналы с геометрическим надзором и пространственно-временной памятью, обеспечивая согласованность в течение длительных прогонов. Между тем, модели открытого мира, такие как Wonderland [20], WonderWorld [51], Wonder-Turbo [26], и EvoWorld [40] дополнительно интегрируют геометрически индексированные или адаптивные памяти для поддержания постоянных состояний мира в ходе взаимодействий. Однако эти подходы все еще используют неявные, ограниченные клипами памяти, тогда как мы вводим явную поза-индексированную мировую память, извлекаемую по запросу для генерации, управляемой камерой.

2.2. Генерация видео с управлением камерой

Ранние модели T2V/I2V изучали движение камеры неявно и испытывают трудности с надежным повторением явных траекторий [11, 42, 61]. Недавние работы, такие как CameraCtrl [10], рассматривают параметры камеры как явные условия, кодируя внешние параметры камеры и траектории или применяя ограничения пути — для улучшения управляемости и точности [2, 25, 41, 47, 55]. Motion-Prompting [9] реализует ко-мпозиционное управление через условие на точечные треки, а MotionPro [56] использует потери выравнивания пути, которые снижают ошибки вращения и трансляции; управление без обучения также достигается путем подгонки легкого облака точек и использованием предварительного шума для управления денойзингом [11]. Геометрические приоритеты, сохраняющие сцену, дополнительно укрепляют согласованность на уровне клипа. Cam2IV [57] рассматривает позу камеры как физический приоритет и использует эпиполярные и многовидовые ограничения; RealCam-I2V [19] восстанавливает метрическую глубину с помощью DepthAnything v2 [48] для реконструкции сцены с устойчивым масштабом; PoseTraj [16] использует предварительное обучение, учитывающее позу, для получения

движения, выровненного по вращению. По сравнению с условием только на параметры, эти приоритеты уменьшают дрейф компоновки внутри клипа и лучше сохраняют локальную геометрию при изменении вида. Более того, недавние работы связывают управление камерой с моделированием мира. CVD [17], Cavia [46], и WoVoGen [22] совместно синтезируют многовидовые и многотраекторные видео из общей презентации сцены, обеспечивая согласованность между путями. Между тем, методы, которые используют условие на явныерендерируемые 3D представления (например, 3D Гауссианы), могут закреплять геометрию, улучшать 3D согласованность между видами и соблюдение пути [15, 29, 33, 52, 53]. Однако эти подходы обычно строят одноразовые 3D сцены, тогда как мы объединяем управление камерой на длинной дистанции с постоянной памятью мира, индексированной по позе, общкой для всех траекторий.

3. Captain Safari

Мы представляем *Captain Safari*, структуру генерации видео, управляемую памятью. Раздел 3.1 представляет собой неявную память мира для стабильного представления сцены, в то время как раздел 3.2 описывает систему извлечения, обусловленную позой, которая сопоставляет виды камеры с токенами мира, направляя генератор на основе DiT для получения согласованных результатов вдоль произвольных траекторий.

3.1. Неявная память геометрии мира

у

Постановка задачи. Мы представляем видео как $V = \{I_t\}_{t=0}^T$, где I_t — это кадр на временном шаге t . На той же временной оси мы определяем позы камеры $\mathcal{C} = \{(R_t, T_t)\}_{t=0}^T$ и получаем 3D-осведомленную характеристику памяти m_t на каждом временном шаге t с использованием предварительно обученного геометрического кодировщика. Все характеристики памяти формируют глобальный банк памяти $\mathcal{M} = \{m_t\}_{t=0}^T$.

Имея текстовый запрос p , позы камеры \mathcal{C} и цель

шаг времени клипа $\mathcal{T} = [t_0, t_1]$, вместе с его локальной памятью мира $\mathcal{M}_{\text{local}} \subset \mathcal{M}$, наша цель — синтезировать видеосегмент $V_{\mathcal{T}}$, который (i) соответствует p , (ii) соблюдает предписанные позы $\{(R_t, T_t)\}_{t \in \mathcal{T}}$, и (iii) поддерживает целостный 3D мир между точками обзора.

Локальная память мира. Прямое использование полного банка памяти \mathcal{M} для каждого клипа было бы вычислительно затратным и подвержено влиянию временно удаленных наблюдений. Вместо этого для каждого целевого шага времени клипа $\mathcal{T} = [t_0, t_1]$ мы определяем локальную память $\mathcal{M}_{\text{local}} = \{m_{\tau} \mid \tau \in [k_s, k_e]\}$, конечные точки которой выбираются

$$\begin{aligned} t_0 - L &\leq k_s \leq t_0, \\ \max(k_s, t_0) + 1 &\leq k_e \leq \min(k_s + L, t_1), \end{aligned} \quad (1)$$

где L — фиксированная граница, а все шаги времени — целые числа. Эти ограничения обеспечивают, что: (i) окно памяти начинается не более чем за L секунд до входа в клип t_0 , связывая его с близлежащими наблюдениями; (ii) его продолжительность составляет не более L , что сохраняет компактность набора условий; и (iii) его конечное время k_e всегда касается или перекрывает t_0 , оставаясь в пределах $[t_0, t_1]$, обеспечивая поддержку каждого клипа временно совместимым мировым приоритетом. Все $\mathcal{M}_{\text{local}}$ строятся как такие динамические окна, выровненные по клипам, общего банка \mathcal{M} , так что соседние клипы естественным образом разделяют перекрывающиеся записи памяти, ограничивая вычисления и связывая их генерации с 3D-согласованным основным миром.

Память, извлеченная по позе. В рамках заданного шага времени клипа \mathcal{T} , мы рассматриваем локальную память $\mathcal{M}_{\text{local}}$ как статическую гипотезу окружающего мира, построенную из ключевых кадров. Каждый шаг времени τ предоставляет токен позы p_{τ} (полученный из (R_{τ}, T_{τ})) и набор 3D-осведомленных токенов памяти $m_{\tau,1}, \dots, m_{\tau,M}$. Коллекция $\{(p_{\tau}, m_{\tau,1:M})\}_{\tau}$ формирует неявную таблицу мира: токен позы указывает, где камера наблюдала сцену, в то время как токены памяти кодируют, как мир выглядит из этих конфигураций. Для любого целевого шага времени $t \in \mathcal{T}$, мы выводим его позу камеры в токен запроса позы p_t , встраиваем его как $q_t = \phi_p(p_t)$, и используем специальный модуль извлечения для чтения из этой статической таблицы в зависимости от позы. Конкретно, q_t конкatenируется с банком обучаемых токенов запроса и обрабатывается в запросы извлечения, которые выполняют перекрестное внимание над закодированной памятью X^{mem} (определенной в Разделе 3.2), получая набор токенов мира.

$$w_t = \text{Agg}\left(\text{CrossAttn}(Q_t, \tilde{X}^{\text{mem}})\right). \quad (2)$$

соответствующие обновленным обучаемым запросам. Эти выровненные по позе токены мира w_t используются непосредственно как реконструированная память в позе t . Таким образом, все кадры в \mathcal{T} доступ к локальной памяти осуществляется через запросы, обусловленные позой, вместо использования сырых временных индексов, что способствует сохранению многовидовых наблюдений, связанных с согласованным статическим 3D миром.

3.2. Извлечение и кондиционирование памяти

Дизайн извлекателя памяти. Как показано на Рисунке 2, учитывая локальную память, мы представляем каждый временной шаг τ позой

токен p_{τ} и его связанные токены памяти $m_{\tau,1:M}$. Наш извлекатель разработан для того, чтобы (i) совместно кодировать пары позы-память в целостное представление мира, и (ii) извлекать для любого запроса позы компактный набор токенов, выровненных по позе, которые суммируют наиболее релевантные части этого локального мира.

Сначала мы встраиваем особенности позы и памяти в общее пространство и формируем совместную последовательность для каждого временного шага:

$$\hat{X}_{\tau} = [\phi_p(p_{\tau}), \phi_m(m_{\tau,1}), \dots, \phi_m(m_{\tau,M})], \quad (3)$$

где ϕ_p и ϕ_m обозначают обучаемые встраивания для токенов позы и памяти соответственно. Стек трансформерных блоков (MemEnc) с 3D-осведомленным позиционным кодированием уточняет эти последовательности,

$$\tilde{X}_{\tau} = \text{MemEnc}(\hat{X}_{\tau}), \quad (4)$$

и мы получаем закодированную локальную память мира путем конкатенации

$$\tilde{X}^{\text{mem}} = [\tilde{X}_{k_s}, \dots, \tilde{X}_{k_e}], \quad (5)$$

с возможным маскированием для исключения заполненных или неключевых записей.

Для целевого временного шага t мы выводим токен запроса позы p_t , встраиваем его как $q_t = \phi_p(p_t)$ и конкatenируем с M обучаемыми токенами запроса r_1, \dots, r_M ,

$$\hat{Q}_t = [q_t, r_1, \dots, r_M]. \quad (6)$$

Эта последовательность уточняется трансформерными блоками, имеющими ту же архитектуру, что и MemEnc, обозначенными как QryEnc, создавая запросы извлечения, учитывающие позу.

$$Q_t = \text{QryEnc}(\hat{Q}_t). \quad (7)$$

Затем мы выполняем перекрестное внимание от Q_t к закодированной памяти \tilde{X}^{mem} ,

$$Y_t = Q_t + \text{CrossAttn}(Q_t, \tilde{X}^{\text{mem}}), \quad (8)$$

и берем подмножество токенов в Y_t , соответствующих обучаемым запросам, как извлеченные токены мира

$$w_t = [w_{t,1}, \dots, w_{t,M}], \quad (9)$$

которые формируют выровненную по позе мировую характеристику для времени t . Во время обучения линейная голова отображает w_t обратно в исходное пространство памяти, чтобы реконструировать целевые токены памяти в позе запроса. Многократное наложение блоков извлечения итеративно уточняет как запросы, так и извлеченные токены, позволяя модели мягко направлять каждый запрос позы к наиболее релевантному подмножеству прошлых наблюдений, вместо того чтобы полагаться на жесткое временное соседство или единственный ближайший кадр.

DiT с памятью. Для заданного целевого шага времени клипа \mathcal{T} , извлекатель обрабатывает $\mathcal{M}_{\text{local}}$ и позу запроса p_t и выдает набор токенов мира $w_t \in \mathbb{R}^{M \times d_m}$, выровненных по позе, которые суммируют статический локальный мир, относящийся к этому сегменту. Эти токены отображаются в скрытое пространство DiT с помощью MLP для встраивания памяти.

$$W_{\mathcal{T}} = \phi_w(w_t) \in \mathbb{R}^{M \times D}. \quad (10)$$

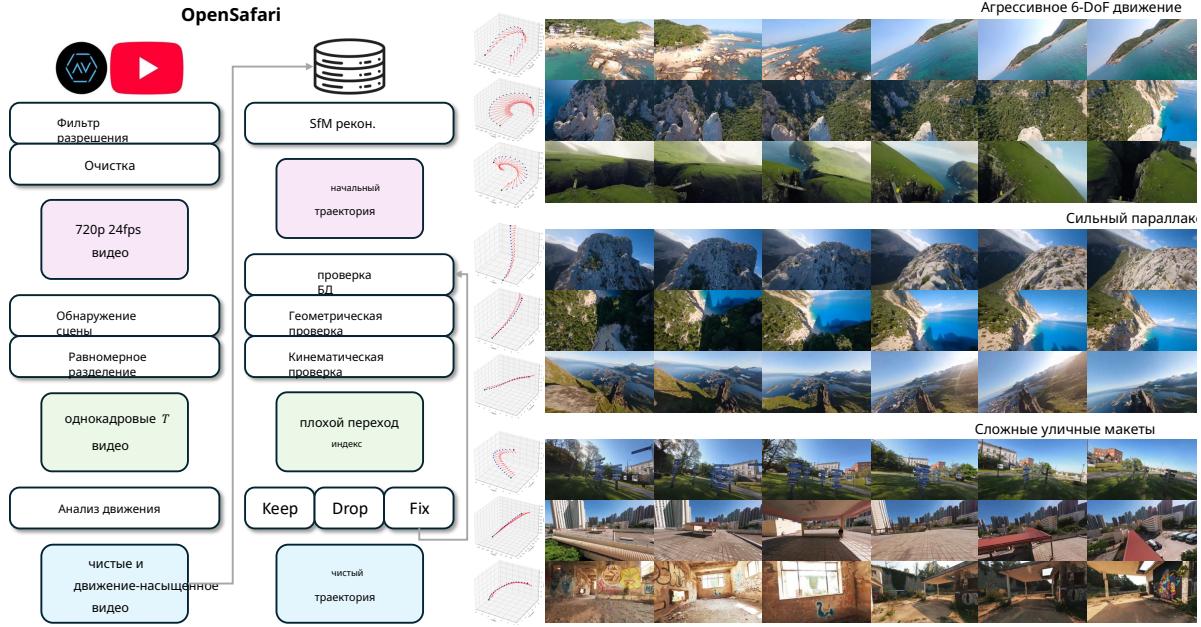


Рисунок 3. *OpenSafari*. Новый FPV набор данных в естественных условиях с тщательно проверенными траекториями камеры, разработанный для стресс-тестирования геометрически согласованной, управляемой камерой генерации видео. Мы курируем клипы через компактный, многоэтапный конвейер, который фильтрует, реконструирует и проверяет траектории, создавая чистые, насыщенные движением видео с надежными путями камеры.

Латентный клип кодируется как единая последовательность пространственно-временных токенов $Z \in \mathbb{R}^{L_z \times D}$, полученная путем разбиения на пачки всех кадров в V_T . На каждом слое DiT мы сначала применяем самовнимание ко всей последовательности, а затем вводим токены мира через специальное перекрестное внимание памяти:

$$Z^{(l+1)} = Z^l + \text{CrossAttn}(Z^l, W_{\mathcal{T}}, W_{\mathcal{T}}). \quad (11)$$

Клип-уровневые токены мира $W_{\mathcal{T}}$ используются повторно в качестве ключей и значений на всех уровнях, обеспечивая стабильный, 3D-согласованный приоритет, который формирует удаление шума каждого пространственно-временного токена.

4. OpenSafari

4.1. Куратория видео данных

Существующие наборы данных, зависящие от камеры, не соответствуют нашему целевому режиму. RealEstate10K [59] сосредоточен на медленных, в основном внутренних турах по недвижимости с плавным движением и чистыми, квазистатичными сценами, в то время как MineCraft [4] является синтетическим воксельным миром с упрощенной геометрией и динамикой, ограниченной движком. Ни один из них не захватывает агрессивные, в дикой природе полеты дронов с 6-DoF, сильным параллаксом, значительными изменениями высоты и сложными уличными макетами, которые действительно испытывают долгосрочную 3D согласованность. Поэтому мы предлагаем *OpenSafari*, новый набор данных реальных видео с дронов в стиле FPV с проверенными траекториями камеры, адаптированными к этой сложной обстановке.

Мы создаем Safari-FPV из видео в стиле FPV-дронов

собранных на AirVuz¹ и YouTube², и сохраняем только клипы, которые проходят строгий многоэтапный процесс предварительной обработки. Как показано на Рисунке 3, мы: (i) загружаем наивысшее доступное разрешение для каждого URL и отбрасываем источники ниже целевого разрешения; (ii) нормализуем все видео до 720p, 24fps и фиксированного 16:9 центрального кадрирования, удаляя черные полосы и границы, чтобы последующая оценка камеры выполнялась на чистом поле зрения; (iii) запускаем обнаружение сцен для получения одноразовых сегментов; (iv) разделяем сегменты на видео фиксированной длины с помощью равномерного временного нарезания.

Затем мы фильтруем видео с помощью одного диагностического метода, основанного на движении. В частности, мы запускаем RAFT [36] для оценки величины оптического потока; видео со недостаточным движением удаляются, в то время как видео со стабильным, согласованным движением сохраняются, чтобы подчеркнуть информативные, насыщенные параллаксом траектории, а не статичные виды. Только видео, удовлетворяющие условию движения, попадают в окончательный набор данных. Это приводит к созданию крупномасштабного корпуса видео с дронов в дикой природе, специально предназначенного для стресс-тестирования генерации видео, учитывающей геометрию и следование траектории.

4.2. Восстановление траектории камеры

Для каждого отобранного Видео мы оцениваем внутренние и внешние параметры камеры с частотой 4 кадра в секунду, используя Иерархическую локализацию [30, 31]. Мы извлекаем локальные признаки, создаем исчерпывающие пары изображений в каждом Видео, выполняем сопоставление признаков и реконструируем модель SfM в стиле COLMAP; из этой модели мы экспортируем

¹<https://www.airvuz.com/>
²<https://www.youtube.com/>

Таблица 1. Бенчмарк генерации видео с управлением камерой. *Captain Safari* занимает первое место по 3D согласованности и следованию траектории с конкурентоспособным качеством видео. По сравнению с вариантом без памяти, *Captain Safari* значительно улучшает 3D согласованность и следование траектории, с лишь небольшим компромиссом в качестве видео. (Рекон. = скорость реконструкции. CosSim = косинусное сходство.)

Модель	Качество видео		3D согласованность		Следование траектории		
	FVD ↓	LPIPS ↓	MEt3R ↓	Рекон. ↑	AUC@30 ↑	AUC@15 ↑	КосСим ↑
Geometry Forcing [44]	2662.75	0.667	0.4834	0.877	0.168	0.056	0.429
Real-CamI2V [19]	1585.61	0.513	0.3703	0.923	0.174	0.051	0.296
Wan2.2-5B-Control-Camera [38]	1387.75	0.545	0.3932	0.767	0.181	0.054	0.420
Captain Safari без памяти	998.47	0.504	0.3720	0.912	0.193	0.068	0.508
Captain Safari	1023.46	0.512	0.3690	0.968	0.200	0.068	0.563

Таблица 2. Предпочтения пользователей. Пользователи в подавляющем большинстве предпочитают *Captain Safari* по всем критериям, что составляет 67% от общего числа голосов. Вариант без памяти занимает далеское второе место, в то время как базовые модели получают однозначные предпочтения.

Модель	Качество видео	3D согласованность	Следование траектории	Среднее
Geometry Forcing [44]	0,20%	0,00%	0,20%	0,13%
Real-CamI2V [19]	4,20%	6,40%	4,40%	5,00%
Wan2.2-5B-Control-Camera [38]	3,20%	3,80%	6,40%	4,47%
Captain Safari без памяти	25,00%	24,20%	20,00%	23,07%
Captain Safari	67,40%	65,60%	69,00%	67,33%

параметры камеры для каждого кадра в качестве начальных траекторий.

Чтобы получить данные, готовые к развертыванию, мы применяем трехэтапный процесс проверки и исправления к каждой восстановленной траектории. Сначала проверка базы данных использует статистику SfM (количество и соотношение инлайперов) для выявления потенциально ненадежных переходов. Затем геометрическая проверка пересматривает подозрительные пары, используя сохраненные ключевые точки и совпадения, пересчитывает основные матрицы и устанавливает пороги симметричных эпиполярных ошибок. Наконец, кинематическая проверка анализирует последовательность поз на наличие всплесков трансляции, скачков вращения, переворотов в направлении движения и нарушений плавности более высокого порядка, используя надежные оценки на основе MAD для обнаружения неправдоподобного движения.

Решения по каждому переходу объединяются в двоичный индекс плохих переходов, который определяет строгую политику. Если плохие переходы редки и локализованы, мы применяем целенаправленное исправление: линейно интерполируем центры камер и применяем SLERP к вращениям с ограниченным углом интерполяции, при необходимости экстраполируя на границах видео.

Исправленные сегменты затем повторно проверяются по тем же критериям базы данных/геометрии/кинематики. Если проверка после исправления успешна, траектория экспортируется в окончательный набор данных. Если индекс плохих переходов слишком плотный, нарушения слишком серьезны или исправленные траектории все еще не проходят проверку, все видео отбрасывается.

Полученный *OpenSafari* сочетает высокодинамичное FPV видео дронов в естественной среде с тщательно проверенными траекториями камеры. Он отличается от существующих бенчмарков, акцентируя внимание на агрессивном движении 6-DoF, сильном параллаксе и сложных уличных планировках, при этом обеспечивая строгую геометрическую и кинематическую валидацию. Это делает *OpenSafari* сложной тестовой платформой для генерации видео с управляемой камерой.

5. Эксперименты

5.1. Детали реализации

Рецепт обучения. Мы используем двухэтапный рецепт. Сначала мы разогреваем поисковик памяти, зависящий от позы, используя токены памяти, выровненные по позе m_t . Затем мы совместно обучаем поисковик и DiT от начала до конца, обновляя DiT через LoRA [12]. Перекрестное внимание к памяти инициализируется из соответствующих весов перекрестного внимания к контексту, а другие новые слои используют стандартную инициализацию.

Набор данных. Мы извлекаем перекрывающиеся клипы с шагом 1с, получая 51,997 кандидатов для обучения. Фильтр траекторий на основе разнообразия удаляет клипы с почти статичным движением, в результате чего остается 11,481 окончательных обучающих клипов. Дополнительно мы создаем неперекрывающийся тестовый набор из 787 клипов для оценки. Для каждого клипа мы генерируем одно описательное заглавие с использованием Qwen2.5-VL-7B [3] и используем его в качестве текстового условия.

Конфигурации и обозначения. Мы генерируем клипы $T = 5$ с частотой 24 кадра в секунду из видео $T = 15$. Позиции камеры и характеристики памяти выбираются с частотой 4 кадра в секунду. Для целевого клипа длительностью 5 секунд с интервалом $[t_0, t_1]$ мы используем конечную позу p_{t_1} в качестве запроса. Окно памяти ограничено $L = 5$ секундами. Мы используем Wan2.2-Fun-5B-Control-Camera [38] в качестве нашей базовой DiT с скрытым размером $D = 3072$. Извлекатель и DiT обучаются на 1 и 5 эпохах соответственно. Для каждого в идею мы извлекаем 3D-осведомленную характеристику памяти из предварительно обученной StreamVG-GT [62]. Мы выбираем четыре слоя $\{4, 11, 17, 23\}$; на каждом слое характеристика содержит 782 токена. Конкатенация через четыре слоя дает $M = 4 \times 782$ и $d_m = 1024$ токенов памяти на кадр.

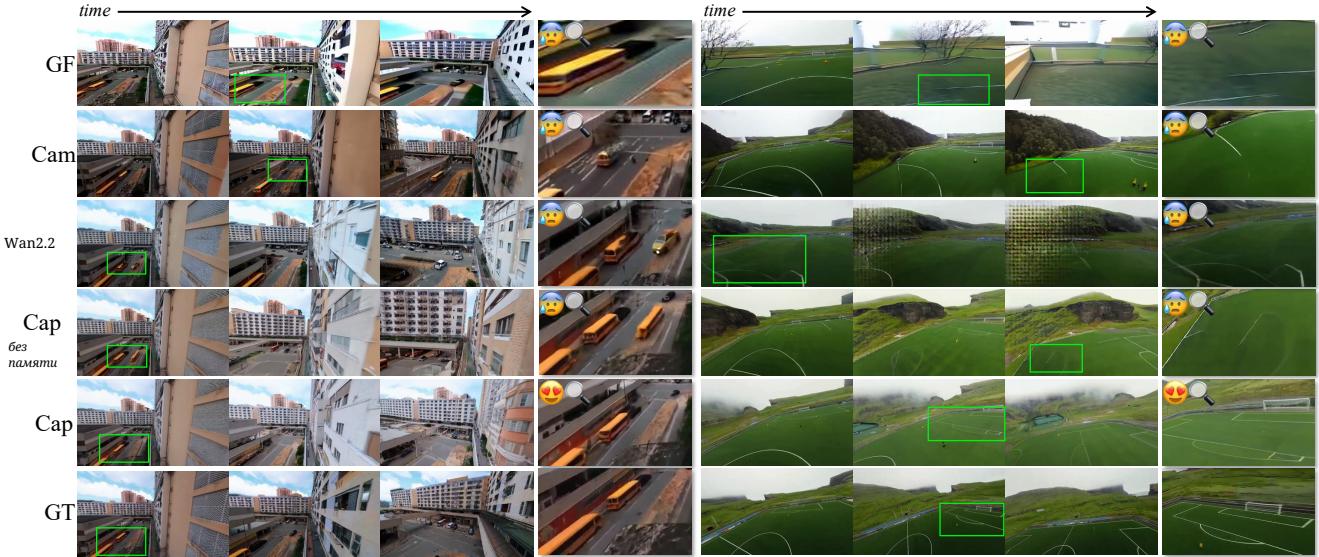


Рисунок 4. Качественные сравнения. Слева: Базовые модели — включая вариант без памяти — демонстрируют резкое появление/исчезновение школьного автобуса, а GF низкого качества. Только *Captain Safari* плавно выводит автобус из кадра. Справа: Базовые модели искажают или теряют разметку поля, при этом Wan2.2 рушится при большом движении камеры, подтверждая сложность 3D согласованности при быстрых траекториях. *Captain Safari* сохраняет четкие разметки и согласованную компоновку, следя по быстрому пути 6-DoF.

5.2. Бенчмарк

Метрики. Мы оцениваем генерацию видео по трем дополнительным осям: качество видео, 3D согласованность и следование траектории. Для качества видео мы сообщаем FVD [37] и LPIPS [54]. Для 3D согласованности мы используем MEt3R [1], вычисленный между GT и сгенерированными видео на совпадающих временных шагах, и коэффициент рекоинструкции, который измеряет процент кадров, успешно зарегистрированных в восстановленной 3D модели [30, 31]. Для следования траектории мы сообщаем точность перемещения камеры (AUC [39]) и косинусное сходство между уплощенной позой камеры, показывающее, как модель придерживается желаемых параметров камеры со временем.

Базовые модели. Мы сравниваем с представительными моделями генерации видео с управляемой камерой, включая Geome-try Forcing [44], Real-CamI2V [19, 57], и Wan2.2-5B-Control-Camera [38], которые охватывают подходы, основанные на геометрических ограничениях, реконструкции и крупномасштабной диффузии для синтеза видео, обусловленного траекторией.

Исследование с участием людей. Мы проводим исследование с участием 50 участников. Каждому участнику представлено 10 случаев, где каждый случай содержит GT-видео и пять анонимных видео, сгенерированных моделями (три базовые модели, наша модель и ее вариант без некоторых компонентов). В каждом случае участникам предлагается выбрать лучшее видео по трем критериям: качество видео, 3D согласованность и следование траектории. В общей сложности исследование собирает $50 \times 10 \times 3 = 1,500$ голосов предпочтений участников.

5.3. Качество генерации

Как показано в Таблице 1, наш *CaptainSafari* достигает значительно более низкого FVD (1023.46 против 1387.75) и немного

улучшенного показателя LPIPS (0.512 против 0.513) по сравнению с базовым уровнем SOTA, демонстрируя более стабильную временную динамику и более четкие пространственные детали. Более того, исследование с участием людей в Таблице 2 показывает, что **67.40%** участников предпочитают наши видео по сравнению с конкуртирующими методами, что подчеркивает перцептивный реализм и общую достоверность наших поколений.

Качественные сравнения на Рисунке 4 дополнительно показывают, что *Captain Safari* создает визуально привлекательную, реалистичную и высоко аутентичную динамику сцен. Эти результаты также согласуются с образцами, представленными на Рисунке 1, где наш метод обеспечивает яркие, последовательные и естественно выглядящие дрон-видео, которые близко напоминают реальные съемки.

5.4. 3D согласованность

Captain Safari достигает передовой 3D согласованности. Как показано в Таблице 1, наш метод снижает MEt3R на 0.0013 (0.3690 против 0.3703) и увеличивает скорость реконструкции на 0.045 (0.968 против 0.923) по сравнению с самой сильной базовой линией. Последовательно, исследование с участием людей в Таблице 2 показывает, что **65.60%** участников предпочитают *Captain Safari* за 3D согласованность, значительно превосходя все конкуртирующие подходы.

Качественные визуализации дополнительно подтверждают эти количественные улучшения. На Рисунке 1 такие структуры, как колонны в греческом стиле, остаются геометрически стабильными при значительных изменениях угла обзора. На Рисунке 4 наша модель создает (слева) школьный автобус, который плавно выходит из кадра, и (справа) сохраняет четкие, глобально согласованные разметки на футбольном поле, тогда как базовые модели демонстрируют искажения и исчезновение. Рисунки 5 и Рисунок 1 дополнительно показывают, что наши реконструкции дают более четкие фасады и хорошо сформированные окна

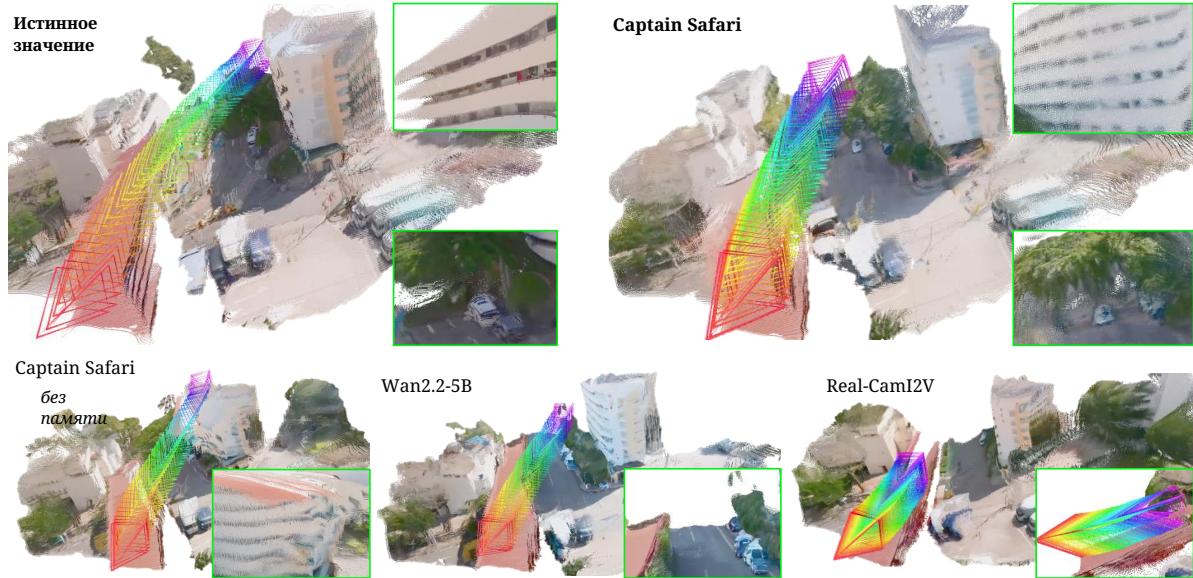


Рисунок 5. Реконструкция сцены и траектория камеры. С использованием памяти, выровненной по позе, *Captain Safari* реконструирует хорошо структурированный фасад здания (вариант без памяти размыает/искажает его), демонстрируя преимущество памяти. Он также сохраняет мелкие детали — припаркованные автомобили и дерево на их крышах — которые Wan2.2-5B не удается удержать. Между тем, Real-CamI2V следует только короткому пути, тогда как *Captain Safari* охватывает всю траекторию со стабильной 3D структурой, подчеркивая сложность поддержания 3D согласованности при быстром движении.

без разрушения геометрии. В совокупности эти результаты подтверждают, что неявная память мира и извлечение, обусловленное позицией, в *CaptainSafari* эффективно стабилизируют подлежащий 3D мир при агрессивном движении камеры.

5.5. Следование траектории

CaptainSafari обеспечивает наиболее точное следование траектории среди всех конкурирующих моделей. Как показано в таблице 1, наш метод достигает наивысших значений AUC@30 (0.200) и AUC@15 (0.068), а также лучшего косинусного сходства (0.563), превосходя сильнейшую базовую линию со явным отрывом. Исследование с участием людей в таблице 2 дополнительно подтверждает это наблюдение, **69.00%** участников определили нашу модель как наиболее точно соответствующую целевой траектории камеры.

Рисунок 5 предоставляет четкую визуализацию этих улучшений. *Captain Safari's* предсказанная траектория тесно совпадает с истинным путем, в то время как вариант с аблацией отклоняется и пролетает над крышей, а RealCam-I2V не удается следовать заданному движению вперед, продвигаясь лишь немногого, вместо того чтобы придерживаться предписанной траектории. Более того, наш метод демонстрирует стабильное и согласованное поколение при сложных изменениях точки зрения с комплексными маневрами камеры на рисунке 1. Эти результаты подчеркивают эффективность нашего дизайна с дополненной памятью и условием позы для точного соблюдения траектории.

5.6. Исследование аблации

Наши результаты подчеркивают важность предложенной памяти мира, обусловленной позой. Как показано в Таблице 1, добавление памяти приводит к значительным улучшениям как в 3D согласованности, так и

следовании траектории. Эти достижения подтверждают, что извлечение мировых признаков, выровненных по позе, в целевом кадре предоставляет модели явное понимание того, как сцена должна выглядеть, обеспечивая стабильную геометрию и точное выравнивание движения.

Качественные сравнения на рисунке 4 и рисунке 5 дополнительно иллюстрируют эти эффекты. С памятью генерированные сцены сохраняют глобальную структуру, поддерживают согласованную геометрию между точками зрения. В отличие от этого, вариант с аблацией часто отклоняется и демонстрирует геометрические несоответствия. Вместе эти результаты подтверждают эффективность нашего дизайна с дополненной памятью в стабилизации базового 3D мира и управлении точным движением камеры.

6. Заключение

Мы представили *Captain Safari*, мировой движок, зависящий от позы, построенный на мировой памяти, который позволяет генерировать видео с дальним охватом и 3D согласованностью при сложных траекториях FPV. Вместе с *OpenSafari*, нашим тщательно отобранным набором данных видео с дронов в естественных условиях с проверенными позами камеры, это устанавливает строгий бенчмарк для управляемой генерации видео. *CaptainSafari* значительно улучшает 3D согласованность и точность траекторий по сравнению с предыдущими методами, сохраняя при этом высокую визуальную точность. Хотя система требует значительных ресурсов для вывода, в будущем будет исследоваться создание мировых движков в реальном времени с легкой памятью и более быстрыми генеративными основами. Мы надеемся, что *Captain Safari* и *OpenSafari* будут стимулировать дальнейшие исследования в области моделей постоянного мира и генерации видео с длинным горизонтом.

Ссылки

- [1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele и Jan Eric Lenssen. Met3d: Измерение многовидовой согласованности в сгенерированных изображениях. В материалах конференции по компьютерному зрению и распознаванию образов, страницы 6034–6044, 2025. 7[2] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Сергей Туляков, Gordon Wetzstein и др. Tc4d: Генерация текста в 4D, обусловленная траекторией. В Европейской конференции по компьютерному зрению, страницы 53–72. Springer, 2024. 3[3] Shuai Bai, Kegui Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang и др. Qwen2. 5-v1 технический отчет. arXiv препринт arXiv:2502.13923, 2025. 6[4] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecco, Brandon Houghton, Raul Sampedro и Jeff Clune. Предобучение видео (vpt): Обучение действовать, наблюдая за неразмечеными онлайн-видео. Advances in Neural Information Processing Systems, 35:24639–24654, 2022. 5[5] Yuanhao Cai, He Zhang, Kai Zhang, Yixun Liang, Mengwei Ren, Fujun Luan, Qing Liu, Soo Ye Kim, Jianming Zhang, Zhifei Zhang и др. Встраивание гауссового размыгивания в диффузионный денойзер для быстрой и масштабируемой однотапной генерации и реконструкции изображений в 3D. arXiv препринт arXiv:2411.14384, 2024. 2[6] Yaru Cao, Zhijian He, Lujia Wang, Wenguan Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han и др. Visdrone-det2021: Результаты вызова по обнаружению объектов на дронах. В материалах Международной конференции IEEE/CVF по компьютерному зрению, страницы 2847–2854, 2021. 2[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng и Yinda Zhang. Matterport3d: Обучение на данных RGB-D в закрытых помещениях. arXiv препринт arXiv:1709.06158, 2017. 1[8] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang и Chuang Gan. Adaworld: Обучение адаптируемым мировым моделям с латентными действиями. arXiv препринт arXiv:2503.18938, 2025. 1, 2[9] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun и др. Подсказка движения: Управление генерацией видео с помощью траекторий движения. В материалах конференции по компьютерному зрению и распознаванию образов, страницы 1–12, 2025. 3[10] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li и Ceyuan Yang. CameraCtrl: Обеспечение управления камерой для генерации видео из текста. arXiv препринт arXiv:2404.02101, 2024. 1, 3[11] Chen Hou и Zhibo Chen. Управление камерой для генерации видео без обучения. arXiv препринт arXiv:2406.10126, 2024. 3[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen и др.
- Lora: Низкоранговая адаптация крупных языковых моделей. ICLR, 1(2):3, 2022. 6[13] Ronghang Hu, Nikhila Ravi, Alexander C Berg и Deepak Pathak. Worldsheet: Обертывание мира в 3D лист для синтеза видов из одного изображения. В материалах Международной конференции IEEE/CVF по компьютерному зрению, страницы 12528–12537, 2021. 1, 2[14] Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He и Li Jiang. Memory Forcing: Пространственно-временная память для согласованной генерации сцен в Minecraft. arXiv препринт arXiv:2510.03198, 2025. 2, 3[15] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo и др. Voyager: Долгосрочная и согласованная с миром видео-диффузия для генерации исследуемых 3D сцен. arXiv препринт arXiv:2506.04225, 2025. 3[16] Longbin Ji, Lei Zhong, Pengfei Wei и Changjian Li. Pose-traj: Управление траекторией с учетом позы в видео-диффузии. В материалах Конференции по компьютерному зрению и распознаванию образов, страницы 22776–22785, 2025. 1, 2, 3[17] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas и Gordon Wetzstein. Совместная видео-диффузия: Согласованная многовидео генерация с управлением камерой. Достижения в области нейронных информационных систем, 37:16240–16271, 2024. 3[18] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas и Gordon Wetzstein. Совместная видео-диффузия: Согласованная многовидео генерация с управлением камерой. Достижения в области нейронных информационных систем, 37:16240–16271, 2024. 2[19] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen и др. RealCam-I2V: Генерация видео из реальных изображений с интерактивным сложным управлением камерой. В материалах Международной конференции IEEE/CVF по компьютерному зрению, страницы 28785–28796, 2025. 1, 3, 6, 7[20] Hanwen Liang, Junli Cao, Vedit Goel, Guocheng Qian, Сергей Королев, Demetri Terzopoulos, Konstantinos N Plataniotis, Сергей Туляков и Jian Ren. Wonderland: Навигация по 3D сценам из одного изображения. В материалах Конференции по компьютерному зрению и распознаванию образов, страницы 798–810, 2025. 1, 3[21] Ruoshi Liu, Rundi Wu, Basile Van Hoornick, Pavel Tokmakov, Сергей Захаров и Carl Vondrick. Zero-1-to-3: Генерация 3D объекта из одного изображения без обучения. В материалах Международной конференции IEEE/CVF по компьютерному зрению, страницы 9298–9309, 2023. 2[22] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, Li Zhang, WoVoGen: Диффузия с учетом объема мира для управляемой генерации сцен вождения с несколькими камерами. В Европейской конференции по компьютерному зрению, страницы 329–345. Springer, 2024. 2, 3[23] Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel Khashabi, Rama Chellappa, Alan Yuille и др. GenEx: Генерация исследуемого мира. arXiv препринт arXiv:2412.09624, 2024. 1, 2

[24] Эндрю Мельник, Михал Люблянац, Конг Лу, Ци Ян, Вэймин Жэнь и Хельге Ригтер. Модели диффузии видео: Обзор. arXiv препринт arXiv:2405.03150, 2024. 2[25] Чонг Мо, Миндэнг Чоа, Синтао Ван, Чжаоян Чжан, Ин Шань и Цзян Чжан. Revideo: Пересоздание видео с управлением движением и содержанием. Advances in Neural Information Processing Systems, 37:18481–18505, 2024. 3[26] Чхаоцзюнь Ни, Сяофэн Ван, Чжэн Чжу, Вайце Ван, Хаоюнь Ли, Гошэн Чжао, Цзе Ли, Вэнькан Цинь, Гуань Хуан и Вэньцзюнь Мэй. Wonderturbo: Генерация интерактивного 3D мира за 0,72 секунды. arXiv препринт arXiv:2504.02261, 2025. 1, 3[27] Ава Пун, Гэри Сан, Цзинкан Ван, Юнь Чэн, Цзэ Ян, Сивабалан Манивасагам, Вэй-Чи Ма и Ракель Урта-сун. Нейронная симуляция освещения для городских сцен. Advances in Neural Information Processing Systems, 36 19326, 2023 19291 2 : – . [28] Сантьхос К Рамакришнан, Аарон Гокаслан, Эрик Вайманс, Александр Максымец, Алекс Клегт, Джон Тернер, Эрик Андерсандер, Войцех Галуба, Эндрю Вестбери, Энджел Икс Чанг и др. Набор данных Habitat-matterport 3d (hm3d): 1000 крупномасштабных 3D окружений для воплощенного ИИ. arXiv препринт arXiv:2109.08238, 2021. 1[29] Сюаньчи Жэнь, Тяньчан Шэнь, Цзяхуэй Хуан, Хуань Лин, Ифань Лу, Мерлин Нимье-Давид, Томас Мюллер, Александер Келлер, Саня Фидлер и Цзюнь Гао. Gen3c: Генерация видео с учетом 3D-информации и точным управлением камерой. В материалах конференции по компьютерному зрению и распознаванию образов, страницы 6121–6132, 2025. 1, 3[30] Поль-Эдуар Сарлин, Сезар Кадена, Роланд Зигварт и Марцин Дымчик. От грубого к точному: Надежная иерархическая локализация в крупном масштабе. В CVPR, 2019. 5, 7[31] Поль-Эдуар Сарлин, Даниэль Детон, Томаш Малисевич и Эндрю Рабинович. SuperGlue: Обучение сопоставлению признаков с использованием графовых нейронных сетей. В CVPR, 2020. 5, 7[32] Манолис Савва, Абхишек Кадиан, Александр Максымец, Или Чжао, Эрик Вайманс, Бхавана Джайн, Джулиан Штрауб, Цзя Ло, Владлен Колтун, Джитендра Малик и др. Habitat: Платформа для исследований воплощенного ИИ. В материалах Международной конференции I EEE/CVF по компьютерному зрению, страницы 9339–9347, 2019. 1[33] Мануэль-Андреас Шнайдер, Лукас Х'оллейн и Маттиас Ниснер. Worldexplorer: К созданию полностью навигационных 3D сцен. arXiv препринт arXiv:2506.01799, 2025. 3[34] Синчэн Шуй, Ханхуэй Дин, Чжэньюань Цинь, Хао Ло, Синцзюнь Ма и Дачэн Тао. Управление свободной формой движения: Управление 6D позами камеры и объектов в генерации видео. В материалах Международной конференции I EEE/CVF по компьютерному зрению, страницы 12449–12458, 2025. 2[35] Шухан Тан, Кельвин Вонг, Шэнъулун Ван, Сивабалан Манивасагам, Мэнье Жэнь и Ракель Уртасун. Scenegen: Обучение генерации реалистичных дорожных сцен. В материалах конференции I EEE/CVF по компьютерному зрению и распознаванию образов, страницы 892–901, 2021. 2[36] Закари Тид и Цзя Дэн. RAFT: Рекуррентные преобразования всех пар полей для оптического потока. В Европейской конференции по компьютерному зрению, страницы 402–419. Springer, 2020. 5

[37] Томас Унтертинер, Съёрд Ван Стенкисте, Кароль Курач, Рафаэль Маринье, Марчин Михальски и Сильвен Джелли. К точным генеративным моделям видео: новая метрика и вызовы. arXiv препринт arXiv:1812.01717, 2018. 7[38] Команда Wan, Анг Ван, Баоли Ай, Бин Вэн, ЧАОЦЗЕ Мао, Чен-Вэй Се, Ди Чен, Фэйву Ю, Хаймин Чжао, Цзяньчжо Ян, Цзяньюань Цзян, Цзяюань Ван, Цзинфин Чжан, Цзинхэй Чжоу, Цзинхан Ван, Цзисюань Чен, Кай Чжу, Кан Чжао, Кю Ян, Ляньхуа Хуан, Мэнсян Фэн, Ниньчжан Чжан, Пандон Ли, Пиньлюй У, Жуйхан Чу, Жуйий Фэн, Шивэй Чжан, Сиян Сун, Тао Фан, Тяньсин Ван, Тяньчжоу Гуй, Тиньчжоу Вэн, Тун Шэнь, Вэй Лин, Вэй Ван, Вэй Ван, Воньмон Чжоу, Вэнхе Ван, Вэнтин Шэнь, Вэньюань Ю, Сянъчжун Ши, Сяомин Хуан, Синь Сюй, Янь Коу, Яньчжо Лю, Ифэй Ли, Ицзин Лю, Имин Ван, Иньчжан Чжан, Итун Хуан, Юн Ли, Ю У, Ю Ло, Юлинь Пан, Юньчжэн, Юньтао Хун, Юпэн Ши, Ютун Фэн, Цзэйиньцы Цзян, Чжэн Хань, Чжи-Фань У и Цзыюй Лю. Wan: Открытые и продвинутые крупномасштабные генеративные модели видео. arXiv препринт arXiv:2503.20314, 2025. 6, 7[39] Цзяньюань Ван, Минхao Чен, Никита Караваев, Андреа Ведальди, Кристиан Руппрехт и Дэвид Новотны. Vggt: Трансформер, основанный на визуальной геометрии. В материалах конференции по компьютерному зрению и распознаванию образов, страницы 5294–5306, 2025. 7[40] Цзяхоа Ван, Луоксин Е, Тайминг Лу, Цзюньфэй Сю, Цзяхань Чжан, Юсиин Го, Сидюнь Лю, Рама Челлаппа, Чэн Пэн, Аллан Юилл и др. EvoWorld: Эволюционирующее панорамное мировое поколение с явной 3D памятью. arXiv препринт arXiv:2510.01183, 2025. 1, 2, 3[41] Сян Ван, Ханьцзе Юань, Шивэй Чжан, Дайоу Чен, Цзюниу Ван, Иньчжан Чжан, Юцзюнь Шэнь, Дэли Чжао и Цзинхэй Чжоу. VideoComposer: Композиционный синтез видео с управляемостью движением. Advances in Neural Information Processing Systems, 36:7594–7611, 2023. 3[42] Чжоуся Ван, Цзыянь Юань, Синтао Ван, Яовэй Ли, Тяньшуй Чен, Мэнхан Ся, Пин Ло и Инь Шань. MotionCtrl: Унифицированный и гибкий контроллер движения для генерации видео. В материалах конференции ACM SIGGRAPH 2024, страницы 1–11, 2024. 3[43] Оливия Уайлс, Джорджия Гкиоксари, Ричард Сэлиски и Джастин Джонсон. SynSip: Конечный синтез вида из одного изображения. В материалах IEEE/CVF конференции по компьютерному зрению и распознаванию образов, страницы 7467–7477, 2020. 1, 2 – . [44] Хаою У, Дианкун У, Тянью Хэ, Цзюньлюань Го, Ян Е, Юэзи Дуань и Цзян Бянь. Geometry Forcing: Объединение моделей диффузии видео и 3D представления для согласованного моделирования мира. arXiv препринт arXiv:2507.07982, 2025. 3, 6, 7[45] Цзянцзун У, Лян Хоу, Хаотян Ян, Синь Тао, Е Тянь, Пэнфай Ван, Ди Чжан и Юньхай Тун. VMoba: Смесь блочного внимания для моделей диффузии видео. arXiv препринт arXiv:2506.23858, 2025. 1, 2[46] Дэцзя Сюй, Ифань Цзян, Чен Хуан, Ляньчэн Сун, Торстен Гернот, Лянлян Чоа, Чжаньян Ван и Хао Тан. Cavia: Управляемая камерой многовидовая модель диффузии видео с интегрированным вниманием. arXiv препринт arXiv:2410.10774, 2024. 3

- [47] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang и Arash Vahdat. Camco: Генерация изображений в видео с управлением камерой и 3D-согласованностью. arXiv preprint arXiv:2406.02509, 2024. 3[48] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng и Hengshuang Zhao. Depth anything v2. Достижения в области систем обработки нейронной информации, 37:21875–21911, 2024. 3[49] Xiaoda Yang, Jiayang Xu, Kaixuan Luan, Xinyu Zhan, Hongshun Qiu, Shijun Shi, Hao Li, Shuai Yang, Li Zhang, Checheng Yu и др. Omnicam: Унифицированная мультимодальная генерация видео с управлением камерой. arXiv preprint arXiv:2504.02312, 2025. 2[50] Alex Yu, Vickie Ye, Matthew Tancik и Angjoo Kanazawa. pixelnerf: Нейронные поля излучения из одного или нескольких изображений. В материалах конференции IEEE/CVF по компьютерному зрению и распознаванию образов, страницы 4578–4587, 2021. 1, 2[51] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman и Jiajun Wu. Wonderworld: Интерактивная генерация 3D-сцен из одного изображения. В материалах конференции по компьютерному зрению и распознаванию образов, страницы 5916–5926, 2025. 1, 3[52] Mark YU, Wenbo Hu, Jinbo Xing и Ying Shan. Trajectorycrafter: Перенаправление траектории камеры для монокулярных видео с помощью моделей диффузии. arXiv preprint arXiv:2503.05638, 2025. 3[53] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan и Yonghong Tian. Viewcrafter: Управление моделями диффузии видео для синтеза новых видов с высоким качеством. arXiv preprint arXiv:2409.02048, 2024. 3[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman и Oliver Wang. Неразумная эффективность глубоких признаков как перцепционного метрика. В материалах конференции IEEE по компьютерному зрению и распознаванию образов, страницы 586–595, 2018. 7[55] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin и Weizhi Wang. Tora: Диффузионный трансформер, ориентированный на траекторию, для генерации видео. В материалах конференции по компьютерному зрению и распознаванию образов, страницы 2063–2073, 2025. 3[56] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao и Tao Mei. Motioprop: Точный контроллер движения для генерации изображений в видео. В материалах конференции по компьютерному зрению и распознаванию образов, страницы 27957–27967, 2025. 3[57] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu и Xi Li. Cami2v: Модель диффузии изображений в видео с управлением камерой. arXiv preprint arXiv:2410.15957, 2024. 1, 3, 7[58] Mengqi Zhou, Yuxi Wang, Jun Hou, Shougao Zhang, Yiwei Li, Chuanchen Luo, Junran Peng и Zhaoxiang Zhang. Scenex: Процедурная управляемая генерация крупномасштабных сцен. arXiv preprint arXiv:2403.15698, 2024. 2[59] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe и Noah Snavely. Stereo magnification: Обучение синтезу видов с использованием многоплоскостных изображений. arXiv preprint arXiv:1805.09817, 2018. 2, 5

- [60] Yunsong Zhou, Michael Simon, Zhenghao Peng, Sicheng Mo, Hongzi Zhu, Minyi Guo и Bolei Zhou. Simgen: Генерация сцен вождения с учетом симулятора. Advances in Neural Information Processing Systems, 37:48874, 2024. 2[61] Zhenghong Zhou, Jie An и Jiebo Luo. Latent-reframe: Управление камерой для моделей диффузии видео без обучения. В Proceedings of the IEEE/CVF International Conference on Computer Vision, страницы 12779–12789, 2025. 3[62] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou и Jiwen Lu. Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539, 2025. 6