

Captain Safari: A World Engine

Yu-Cheng Chou¹ Xingrui Wang¹ Yitong Li² Jiahao Wang¹ Hanting Liu¹
Cihang Xie³ Alan Yuille¹ Junfei Xiao^{1✉}

¹Johns Hopkins University ²Tsinghua University ³UC Santa Cruz

<https://johnson111788.github.io/open-safari/>

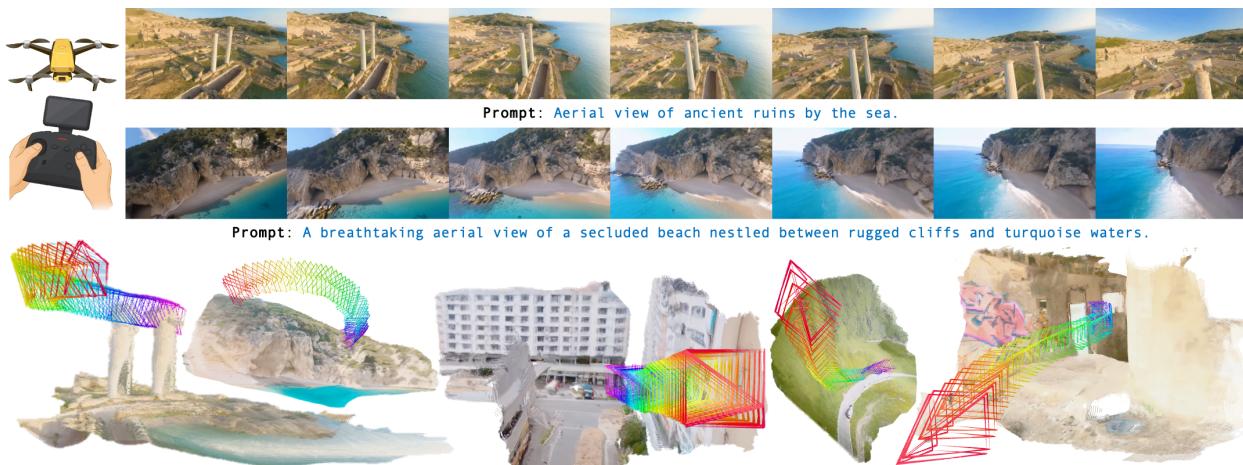


Figure 1. **Captain Safari** is a pose-aware world engine that generates long-horizon, 3D-consistent FPV videos from any user-specified camera trajectory. By retrieving pose-aligned world memory, it keeps geometry stable across large viewpoint changes and reconstructs crisp, well-formed structures while faithfully tracking aggressive 6-DoF motion.

Abstract

World engines aim to synthesize long, 3D-consistent videos that support interactive exploration of a scene under user-controlled camera motion. However, existing systems struggle under aggressive 6-DoF trajectories and complex outdoor layouts: they lose long-range geometric coherence, deviate from the target path, or collapse into overly conservative motion. To this end, we introduce Captain Safari, a pose-conditioned world engine that generates videos by retrieving from a persistent world memory. Given a camera path, our method maintains a dynamic local memory and uses a retriever to fetch pose-aligned world tokens, which then condition video generation along the trajectory. This design enables the model to maintain stable 3D structure while accurately executing challenging camera maneuvers.

To evaluate this setting, we curate OpenSafari, a new *in-the-wild* FPV dataset containing high-dynamic drone videos with verified camera trajectories, constructed through a multi-stage geometric and kinematic validation

pipeline. Across video quality, 3D consistency, and trajectory following, Captain Safari substantially outperforms state-of-the-art camera-controlled generators. It reduces MET3R from 0.3703 to 0.3690, improves AUC@30 from 0.181 to 0.200, and yields substantially lower FVD than all camera-controlled baselines. More importantly, in a 50-participant, 5-way human study where annotators select the best result among five anonymized models, **67.6%** of preferences favor our method across all axes. Our results demonstrate that pose-conditioned world memory is a powerful mechanism for long-horizon, controllable video generation and provide OpenSafari as a challenging new benchmark for future world-engine research.

1. Introduction

Simulating coherent 3D worlds through controllable video generation has long been a foundational challenge for augmented reality, embodied AI, and virtual agents [8–10, 13–16, 19, 20, 23, 26, 29, 40, 43–45, 50, 51, 57]. Classical game engines and physics simulators offer explicit geometry and precise control, but require heavy manual authoring and expensive computation [7, 28, 32]. More-

Captain Safari: Мировой движок

Yu-ChengChou¹ Xingrui Wang¹ Yitong Li² Jiahao Wang¹ Hanting Liu¹CihangXie³ Alan Yuille¹ Junfei Xiao^{1✉}
Университет Джона Хопкинса²Цинхуа университет³Университет Калифорний в
Санта-Крузе<https://johnson111788.github.io/open-safari/>

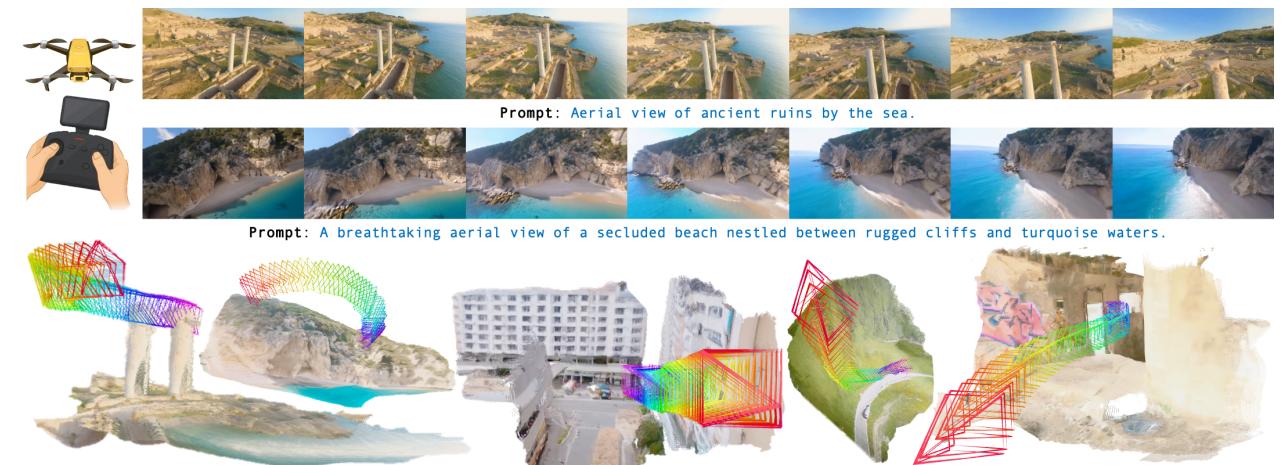


Рисунок 1. CaptainSafari — это мировой движок, учитывающий позу, который генерирует FPV видео с длинным горизонтом и 3D-согласованностью из любой заданной пользователем траектории камеры. Извлекая память мира, согласованную с позой, он сохраняет стабильность геометрии при значительных изменениях точки 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 постянной памяти мира. Учитывая путь камеры, наш метод поддерживает динамическую локальную память и использует извлечатель 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 многоэтапную геометрическую и кинематическую валидацию.

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

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 осмям. 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].

[✉] Corresponding author: Junfei Xiao (xiaojf97@gmail.com)
Preprint, work in progress.

over, while they may achieve visual realism in specialized domains, they still fall short in capturing the richness and diversity characteristic of real world, such as natural scenes [27, 35, 58]. In contrast, modern video diffusion models synthesize high-fidelity, diverse videos from text or images, yet typically operate as feed-forward clip generators without persistent world state: *they struggle with long-range 3D consistency, complex trajectory following, and faithful reconstruction of diverse scenes* [18, 24]. In this work, we move toward bridging this gap with *Captain Safari*, a world engine that enables pose-conditioned modeling of 3D-consistent and diverse environments, surpassing the limitations of traditional game engines in terms of generality, diversity, and interactivity.

Contemporary video world models face three intertwined challenges. First, *long-horizon consistency* is limited by the temporal window of context frames; models often “forget” distant scenery or violate spatial coherence, leading to abrupt appearance changes that break the realism and continuity of the generated environment [8, 14, 45]. Second, achieving *complex camera maneuvers under strict 3D consistency* remains difficult: existing pose- or trajectory-conditioned methods typically work well only for slow, near-forward motions [16, 34, 49]. When the path involves fast 6-DoF movement, strong parallax, or sharp turns, models exhibit a trade-off—either dampening motion and restricting viewpoint changes to preserve geometry, or committing to the requested path at the cost of distortions, flicker, and structural drift. Third, current approaches underrepresent *complex outdoor layouts*. Much of the works focuses on structured, constrained settings (e.g., indoor tours, driving scenes, or real-estate videos), and models are seldom stress-tested in in-the-wild FPV scenarios where the camera weaves around buildings, vegetation, and varied terrain with substantial parallax [6, 22, 59, 60]. As a result, methods that look competitive in simplified environments often fail to preserve geometry and appearance when confronted with truly diverse, complex outdoor scenes.

To address these issues, we introduce *Captain Safari*, a pose-aware world engine that explicitly maintains a persistent notion of world state to uphold *long-horizon 3D consistency* across strong parallax. Because storing and propagating a full long-term state is computationally prohibitive, we develop a retrieval mechanism that *selects and aggregates* only the most informative scene cues, thereby providing strong geometric guidance without incurring prohibitive cost. Crucially, this retrieval is *pose-aware*: given the target camera pose, it assembles a pose-aligned world prior that steers the generation process, enabling accurate tracking of *aggressive camera maneuvers* while preserving 3D-consistent structure in complex environments.

Furthermore, to close the gap in *complex outdoor layouts* and *aggressive camera motion*, we curate *OpenSafari*,

a large-scale dataset of high-dynamic FPV drone videos with verified camera poses. Much of the literature targets structured, constrained settings (e.g., indoor tours, driving or real-estate videos), and even outdoor datasets typically feature slow, near-forward motion. In contrast, *OpenSafari* comprises in-the-wild FPV flights that weave around buildings and vegetation across uneven terrain, exhibiting large parallax, rapid 6-DoF maneuvers, and sharp viewpoint changes. Paired with verified camera trajectories, these videos present diverse, cluttered outdoor scenes and long-range motion, challenging models to maintain 3D consistency while faithfully tracking complex maneuvers.

We evaluate *Captain Safari* along three axes: *video quality*, *3D consistency*, and *trajectory following*. Across these criteria, our method consistently outperforms contemporary camera-controlled video generators on *OpenSafari*: Table 1 reports clear gains in 3D consistency and accurate tracking under complex maneuvers, while maintaining strong perceptual quality. Importantly, a large-scale human study (Table 2) shows that *Captain Safari* receives **67%** of votes in five-way comparisons, indicating that the improvements are perceptually salient. Qualitative comparisons (Fig. 4 and Fig. 5) further demonstrate stable geometry under long-range path and faithful adherence to sharp 6-DoF camera turns in cluttered outdoor scenes.

In summary, our contributions are:

1. We present *Captain Safari*, the first camera-controlled video generation method to enforce long-horizon 3D consistency while tracking aggressive FPV maneuvers.
2. We propose a *pose-guided, long-horizon retrieval* that efficiently reconciles strict 3D consistency with accurate tracking of complex maneuvers.
3. We curate *OpenSafari*, a large-scale in-the-wild FPV dataset with verified camera poses, featuring diverse, cluttered outdoor scenes and rapid 6-DoF motion that stress-test geometry-consistent camera control.
4. In *OpenSafari*, our pose-aware retrieval notably improves video quality, 3D consistency, and trajectory alignment, also achieving a **67%** human preference rate.

2. Related Work

2.1. 3D-Consistent World Models

Early image-to-3D approaches reconstruct geometry indirectly via multi-view consistency or implicit fields, but often fail to maintain coherent structure across large view changes [13, 21, 43, 50]. Recent efforts integrate 3D reasoning into the generative process. DiffusionGS [5] injects Gaussian Splatting into the diffusion denoiser, enforcing view consistency and enabling single-stage, scalable 3D generation. GenEx [23] and EvoWorld [40] extends this idea from static reconstruction to dynamic world creation, generating explorable 360° panoramic environments

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 времененным окном 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 за запрашиваемым путем 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. 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.

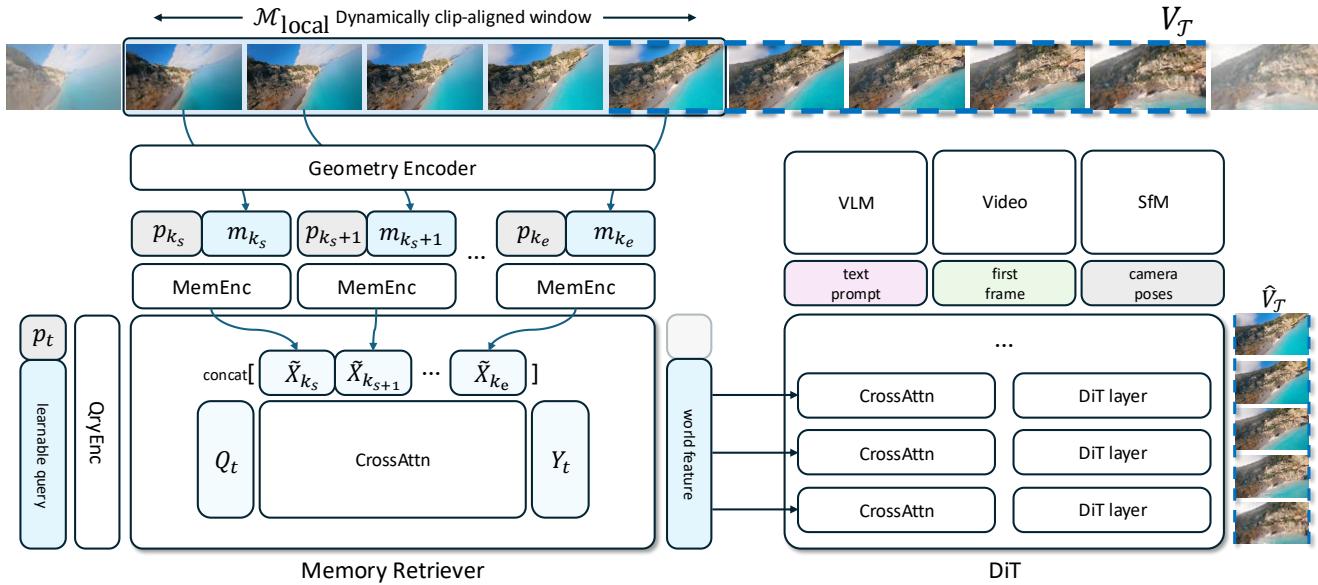


Figure 2. **Method overview.** *Captain Safari* builds a local world memory and, given a query camera pose, retrieves pose-aligned tokens that summarize the scene. These tokens then condition video generation along the user-specified trajectory, preserving a stable 3D layout.

grounded in physical priors. Complementary to these generative reconstructions, Geometry Forcing [44] and Memory Forcing [14] explicitly couple training signals with geometric supervision and spatio-temporal memory, ensuring consistency during long rollouts. Meanwhile, open-world models such as Wonderland [20], WonderWorld [51], Wonder-Turbo [26], and EvoWorld [40] further integrate geometry-indexed or adaptive memories to maintain persistent world states across interactions. However, these approaches still use implicit, clip-bound memories, whereas we introduce an explicit pose-indexed world memory retrieved on demand for camera-controlled generation.

2.2. Camera Controlled Video Generation

Early T2V/I2V models learned camera motion implicitly and struggle to reliably repeat explicit trajectories [11, 42, 61]. Recent work such as CameraCtrl [10] treats camera parameters as explicit conditions, encoding camera extrinsics and trajectories or enforcing path constraints—to improve controllability and accuracy [2, 25, 41, 47, 55]. Motion-Prompting [9] implements compositional control by point-track conditioning and MotionPro [56] use path-alignment losses that lower rotational and translational error; training-free control is also achieved by fitting a lightweight point-cloud and using a noise-layout prior to steer denoising [11]. Scene-preserving geometric priors further strengthen clip-level consistency. Cami2V [57] treats camera pose as a physical prior and exploits epipolar and multiview constraints; RealCam-I2V [19] recovers metric depth with DepthAnything v2 [48] to reconstruct a scale-stable scene; PoseTraj [16] employs pose-aware pretraining to obtain

rotation-aligned motion. Compared with parameter-only conditioning, these priors reduce within-clip layout drift and better preserve local geometry under view changes. Further, recent work links camera control with world modeling. CVD [17], Cavia [46], and WoVoGen [22] jointly synthesize multi-view and multi-trajectory videos from a shared scene representation, enforcing cross-path consistency. Meanwhile, methods that condition from explicit renderable 3D representations (e.g., 3D Gaussians) can anchor geometry, improve cross-view 3D consistency and path adherence [15, 29, 33, 52, 53]. However, these approaches typically build one-off 3D scenes, whereas we unify long-horizon camera control with a persistent pose-indexed world memory shared across trajectories.

3. Captain Safari

We introduce *Captain Safari*, a memory-guided video generation framework. Sec. 3.1 presents an implicit world memory for stable scene representation, while Sec. 3.2 describes a pose-conditioned retrieval system that maps camera views to world tokens, guiding a DiT-based generator for coherent outputs along arbitrary trajectories.

3.1. Implicit Memory of World Geometry

Problem setup. We represent a video as $V = \{I_t\}_{t=0}^T$, where I_t is the frame at time step t . On the same time axis we define camera poses $\mathcal{C} = \{(R_t, T_t)\}_{t=0}^T$ and obtain a 3D-aware memory feature m_t at each time step t using a pretrained geometry encoder. All memory features form a global memory bank $\mathcal{M} = \{m_t\}_{t=0}^T$.

Given a text prompt p , the camera poses \mathcal{C} , and a target

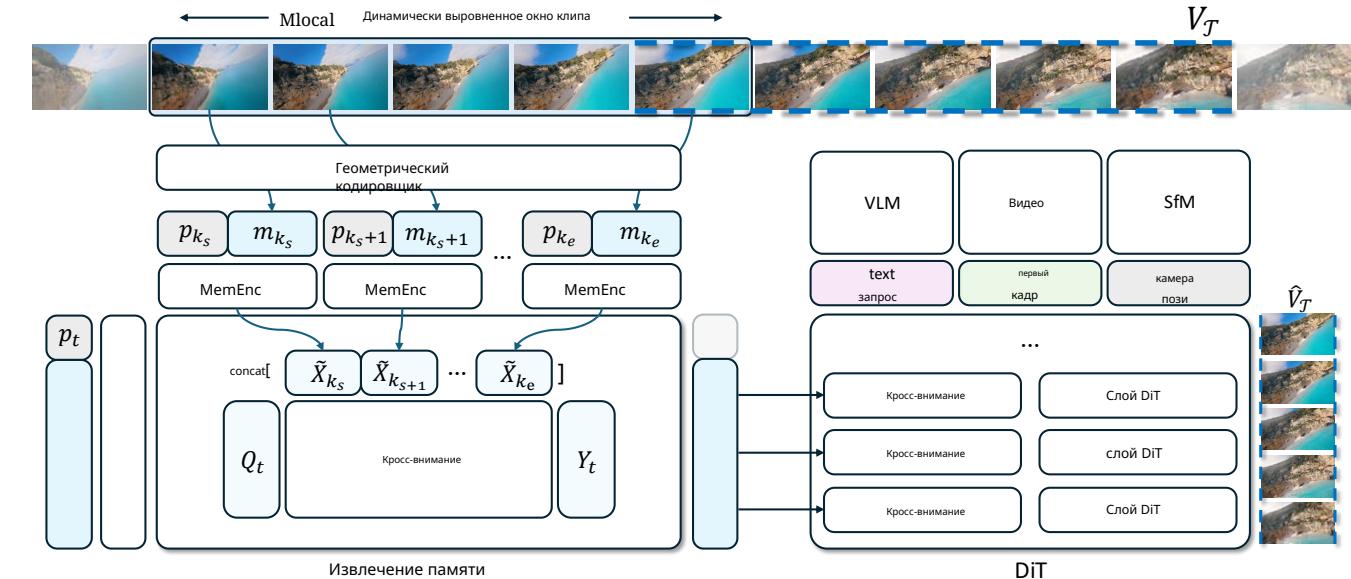


Рисунок 2. Обзор метода. *Captain Safari* создает локальную память мира и, имея заданную позу камеры, извлекает выровненные по позе токены, которые резюмируют сцену. Эти токены затем определяют генерацию видео вдоль заданной пользователем траектории, сохраняя стабильную 3D компоновку.

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 позу-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 композиционное 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 использования предварительного шума для управления денойзингом [11]. Геометрические приоритеты, сохраняющие сцену, дополнительно укрепляют согласованность на уровне клипа. Cami2V [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 явные рендерируемые 3D представления (например, 3D Гауссианы), могут закреплять геометрию, улучшая 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, общей 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

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

clip time step $\mathcal{T} = [t_0, t_1]$, together with its associated local world memory $\mathcal{M}_{\text{local}} \subset \mathcal{M}$, our goal is to synthesize a video segment $\hat{V}_{\mathcal{T}}$ that (i) aligns with p , (ii) respects the prescribed poses $\{(R_t, T_t)\}_{t \in \mathcal{T}}$, and (iii) maintains a coherent 3D world across viewpoints.

Local world memory. Directly conditioning on the full memory bank \mathcal{M} for every clip would be computationally expensive and dominated by temporally distant observations. Instead, for each target clip time step $\mathcal{T} = [t_0, t_1]$ we define a *local* memory $\mathcal{M}_{\text{local}} = \{m_{\tau} \mid \tau \in [k_s, k_e]\}$ whose endpoints are sampled under

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

where L is a fixed bound and all time steps are integers. These constraints enforce that: (i) the memory window starts at most L seconds before the clip entrance t_0 , tying it to nearby observations; (ii) its duration is at most L , which keeps the conditioning set compact; and (iii) its end time k_e always touches or overlaps t_0 while remaining within $[t_0, t_1]$, ensuring that each clip is supported by a temporally compatible world prior. All $\mathcal{M}_{\text{local}}$ are constructed as such dynamic clip-aligned window of the shared bank \mathcal{M} , so neighboring clips naturally share overlapping memory entries, constraining computation while coupling their generations to a 3D-consistent underlying world.

Pose-retrieved memory. Within a given clip time step \mathcal{T} , we treat the local memory $\mathcal{M}_{\text{local}}$ as a static hypothesis of the surrounding world built from key frames. Each time step τ provides a pose token p_{τ} (derived from (R_{τ}, T_{τ})) and a set of 3D-aware memory tokens $m_{\tau,1}, \dots, m_{\tau,M}$. The collection $\{(p_{\tau}, m_{\tau,1:M})\}_{\tau}$ forms an implicit world table: pose token indicates *where* the camera has observed the scene, while memory tokens encode *what* the world looks like from those configurations. For any target time step $t \in \mathcal{T}$, we derive its camera pose to a query pose token p_t , embed it as $q_t = \phi_p(p_t)$, and use a dedicated retrieval module to read from this static table in a pose-dependent manner. Concretely, q_t is concatenated with a bank of learnable query tokens and processed into retrieval queries, which perform cross-attention over the encoded memory \tilde{X}^{mem} (defined in Sec. 3.2), yielding a set of world tokens

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

corresponding to the updated learnable queries. These pose-aligned world tokens w_t are directly used as the reconstructed memory at pose t . Thus, all frames in \mathcal{T} access local memory through pose-conditioned queries instead of raw time indices, encouraging multi-view observations to remain tied to a consistent static 3D world.

3.2. Memory Retrieval and Conditioning

Memory retriever design. As shown in Figure 2, given the local memory, we represent each time step τ by a pose

token p_{τ} and its associated memory tokens $m_{\tau,1:M}$. Our retriever is designed to (i) jointly encode pose–memory pairs into a coherent world representation, and (ii) extract, for any query pose, a compact set of pose-aligned tokens that summarize the most relevant parts of this local world.

We first embed pose and memory features into a shared space and form a joint sequence per time step:

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

where ϕ_p and ϕ_m denote learnable embeddings for pose and memory tokens, respectively. A stack of transformer blocks (MemEnc) with 3D-aware positional encoding refines these sequences,

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

and we obtain the encoded local world memory by concatenation

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

optionally masked to exclude padded or non-key entries.

For a target time step t , we derive the query pose token p_t , embed it as $q_t = \phi_p(p_t)$, and concatenate it with M learnable query tokens r_1, \dots, r_M ,

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

This sequence is refined by transformer blocks sharing the same architecture as MemEnc, denoted as QryEnc, yielding pose-aware retrieval queries

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

We then perform cross-attention from Q_t to the encoded memory \tilde{X}^{mem} ,

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

and take the subset of tokens in Y_t corresponding to the learnable queries as the retrieved world tokens

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

which form a pose-aligned world feature for time t . During training, a linear head maps w_t back to the original memory space to reconstruct the target memory tokens at the query pose. Stacking multiple retrieval blocks iteratively refines both the queries and the retrieved tokens, enabling the model to softly route each query pose to the most relevant subset of past observations, instead of relying on a rigid temporal neighborhood or a single nearest frame.

Memory-conditioned DiT. For a given target clip time step \mathcal{T} , the retriever consumes $\mathcal{M}_{\text{local}}$ and the query pose p_t and outputs a pose-aligned set of world tokens $w_t \in \mathbb{R}^{M \times d_m}$, which summarize the static local world relevant to this segment. These tokens are mapped into the DiT hidden space by the memory embedding MLP

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

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 сырьих временных индексов, что способствует сохранению многовидовых наблюдений, связанных с согласованным статическим 3D миром.

3.2. Извлечение и кондиционирование памяти

Дизайн извлечателя 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 извлечатель разработан для того, чтобы (i) совместно кодировать пары поза–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 конкатенируем 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 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} , извлечатель 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.

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

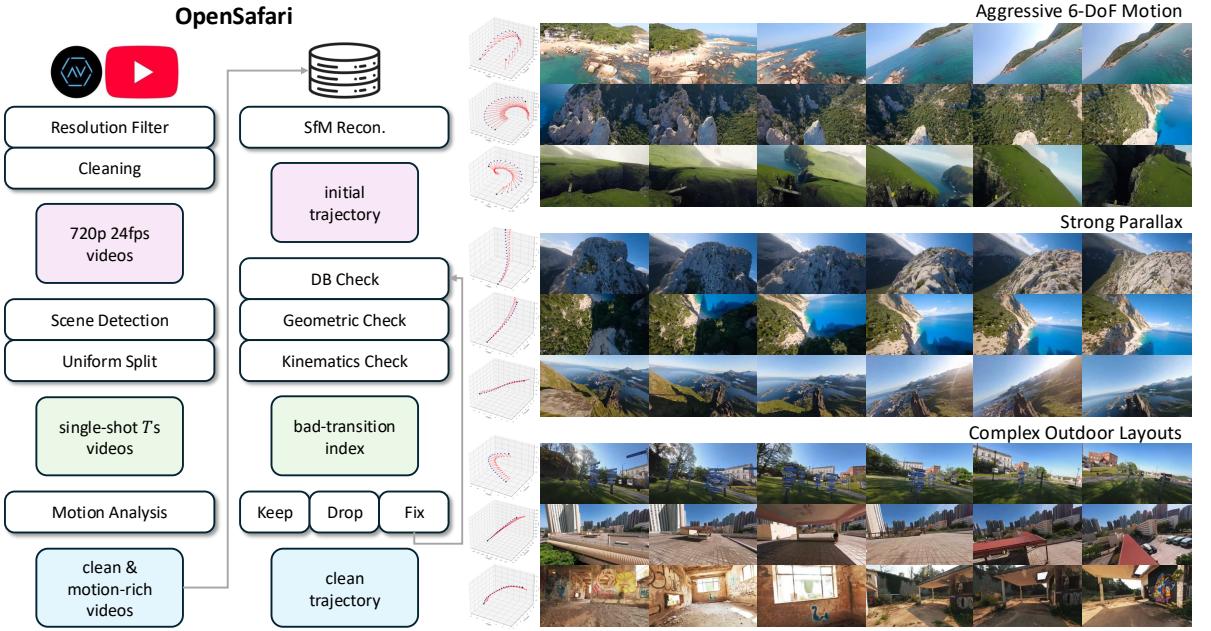


Figure 3. **OpenSafari**. A new in-the-wild FPV dataset with rigorously verified camera trajectories, designed to stress-test geometry-consistent, camera-controllable video generation. We curate clips through a compact, multi-stage pipeline that filters, reconstructs, and verifies trajectories, yielding clean, motion-rich videos with reliable camera paths.

The latent clip is encoded as a single spatio-temporal token sequence $Z \in \mathbb{R}^{L_z \times D}$, obtained by patchifying all frames in V_T . At each DiT layer l , we first apply self-attention over the full sequence and then inject the world tokens through a dedicated memory cross-attention:

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

The clip-level world tokens W_T are reused as keys and values across all layers, providing a stable, 3D-consistent prior that shapes the denoising of every spatio-temporal token.

4. OpenSafari

4.1. Video Data Curation

Existing camera-conditioned datasets do not match our target regime. RealEstate10K [59] focuses on slow, mostly indoor real-estate walkthroughs with gentle motion and clean, quasi-static scenes, while Minecraft [4] is a synthetic voxel world with simplified geometry and engine-constrained dynamics. Neither captures aggressive, in-the-wild 6-DoF drone flight with strong parallax, large elevation changes, and complex outdoor layouts that truly stress long-horizon 3D consistency. We therefore propose *OpenSafari*, a new dataset of real-world FPV-style drone videos with verified camera trajectories tailored to this challenging setting.

We construct Safari-FPV from FPV-style drone videos

collected on AirVuz¹ and YouTube², and retain only clips that pass a strict multi-stage preprocessing pipeline. As shown in Figure 3, we: (i) download the highest available resolution for each URL and discard sources below the target resolution; (ii) normalize all videos to 720p, 24 fps, and a fixed 16:9 center crop, removing letterboxing and black borders so that subsequent camera estimation operates on a clean field of view; (iii) run scene detection to obtain single-shot segments; (iv) split segments into fixed-length T videos via uniform temporal slicing.

We then filter videos with a single diagnostic based on motion. Specifically, we run RAFT [36] to estimate optical-flow magnitude; videos with too little motion are removed, while videos with stable, coherent motion are kept to emphasize informative, parallax-rich trajectories rather than static views. Only videos satisfying the motion constraint enter the final dataset. This yields a large-scale, in-the-wild drone corpus explicitly tailored to stress-test geometry-aware, trajectory-following video generation.

4.2. Camera Trajectory Reconstruction

For each curated video, we estimate camera intrinsics and extrinsics at 4 fps using Hierarchical Localization [30, 31]. We extract local features, build exhaustive image pairs within each video, run feature matching, and reconstruct a COLMAP-style SfM model; from this model we export

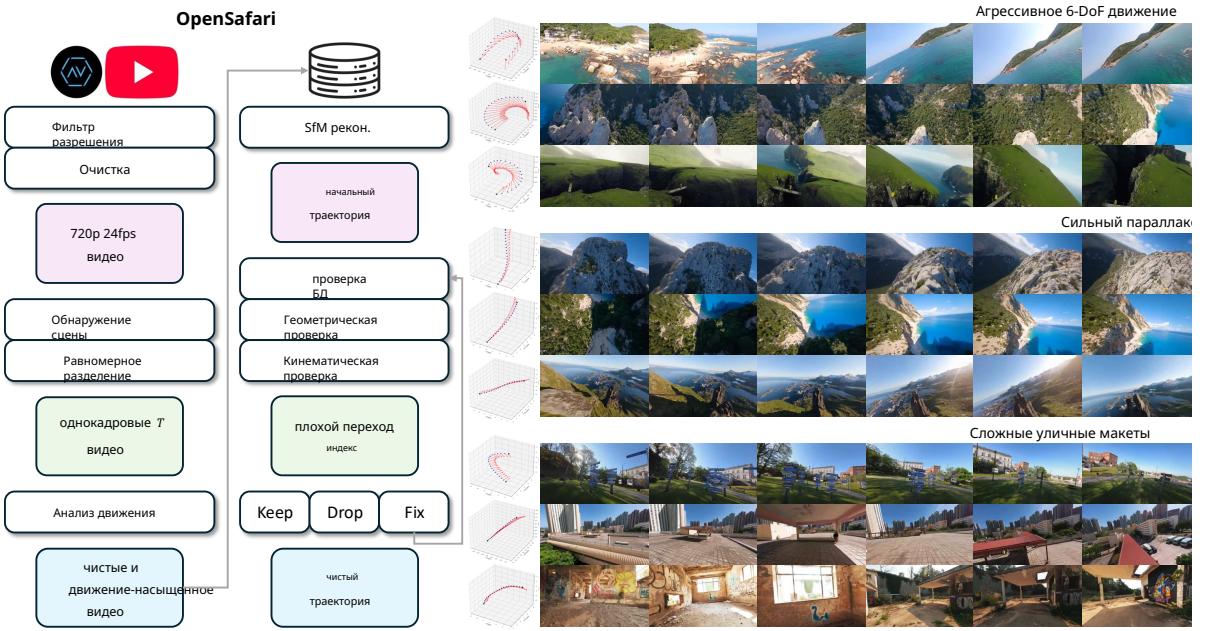


Рисунок 3. **OpenSafari**. Новый FPV набор данных в естественных условиях с тщательно проверенными траекториями камеры, разработанный для стресс-тестирования геометрически согласованной, управляемой камерой генерации видео. Мы курируем клипы через компактный, многоэтапный конвейер, который фильтрует, 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 патчи всех кадров в V_T . На каждом слое DiT l мы сначала применяем самовнимание ко всей последовательности, а затем вводим токены мира через специальное перекрестное внимание памяти:

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

Клип-уровневые токены мира W_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 движением. 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 сцены 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 с 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 условия 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 экспортствуем

¹<https://www.airvuz.com/>

²<https://www.youtube.com/>

Table 1. **Benchmark camera-controlled video generation.** *Captain Safari* ranks first in 3D consistency and trajectory following with competitive video quality. Compared to the ablated variant without memory, *Captain Safari* substantially improves 3D consistency and trajectory following, with only a slight trade-off in video quality. (Recon. = reconstruction rate. CosSim = cosine similarity.)

Model	Video Quality		3D consistency		Trajectory Following		
	FVD ↓	LPIPS ↓	MEt3R ↓	Recon. ↑	AUC@30 ↑	AUC@15 ↑	CosSim ↑
Geometry Forcing [44]	2662.75	0.667	0.4834	0.877	0.168	0.056	0.429
Real-CamI2V [19]	1585.61	0.513	<u>0.3703</u>	<u>0.923</u>	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 w/o Mem.	998.47	0.504	0.3720	0.912	<u>0.193</u>	0.068	<u>0.508</u>
Captain Safari	1023.46	0.512	0.3690	0.968	0.200	0.068	0.563

Table 2. **Human preference.** Users overwhelmingly prefer *Captain Safari* across all criteria, capturing 67% of total votes. The memory-removed variant ranks a distant second, while baselines competitors receive single-digit preference.

Model	Video Quality	3D consistency	Trajectory Following	Average
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 w/o Mem.	25.00%	24.20%	20.00%	23.07%
Captain Safari	67.40%	65.60%	69.00%	67.33%

per-frame camera parameters as initial trajectories.

To obtain deployment-ready data, we apply a three-stage verification-and-fix pipeline to every reconstructed trajectory. First, *database check* consumes SfM statistics (inlier counts and ratios) to flag potentially unreliable transitions. Next, *geometric check* revisits suspicious pairs using stored keypoints and matches, recomputes essential matrices, and thresholds symmetric epipolar errors. Last, *kinematics check* analyzes the pose sequence for translation spikes, rotation jumps, forward-direction flips, and higher-order smoothness violations, using robust MAD-based scores to detect implausible motion.

The per-transition decisions are fused into a binary bad-index, which drives a strict policy. If bad transitions are sparse and localized, we invoke a targeted fix: we linearly interpolate camera centers and apply SLERP to rotations with a capped interpolation angle, optionally extrapolating at video boundaries. The fixed segments are then re-validated by the same database/geometric/kinematics criteria. If post-fix validation succeeds, the trajectory is exported into the final dataset. If the bad-index is too dense, violations are too severe, or fixed trajectories still fail verification, the entire video is discarded.

The resulting *OpenSafari* couples high-dynamic, in-the-wild FPV drone video with rigorously verified camera trajectories. It departs from existing benchmarks by emphasizing aggressive 6-DoF motion, strong parallax, and complex outdoor layouts, while enforcing strict geometric and kinematic validation. This makes *OpenSafari* a challenging testbed for camera-controllable video generation.

5. Experiments

5.1. Implementation Details

Training recipe. We adopt a two-stage recipe. We first warm up the pose-conditioned memory retriever using pose-aligned memory tokens m_t . We then jointly train the retriever and DiT end-to-end, updating the DiT via LoRA [12]. Memory cross-attention is initialized from the corresponding context cross-attention weights, and other new layers use standard initialization.

Dataset. We extract overlapping clips with 1 s stride, yielding 51,997 training candidates. A diversity-based trajectory filter removes clips with near-static motion, resulting in 11,481 final training clips. We additionally construct a non-overlapping test set of 787 clips for evaluation. For each clip, we generate a single descriptive caption using Qwen2.5-VL-7B [3] and use it as the text condition.

Configuration and notation. We generate $\mathcal{T} = 5$ s clips at 24 fps from $T = 15$ s videos. Camera poses and memory features are sampled at 4 fps. For a target 5 s clip with interval $[t_0, t_1]$, we use the terminal pose p_{t_1} as the query. The memory window is limited to $L = 5$ s. We use Wan2.2-Fun-5B-Control-Camera [38] as our base DiT with a hidden dimension $D = 3072$. Retriever and DiT are trained with 1 and 5 epochs, respectively. For each video, we extract 3D-aware memory feature from a pretrained StreamVGTT [62]. We select four layers $\{4, 11, 17, 23\}$; at each layer, the feature contains 782 tokens. Concatenating across the four layers yields $M = 4 \times 782$ and $d_m = 1024$ memory tokens per frame.

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 ↑	KosSim ↑
Geometry Forcing [44]	2662.75	0.667	0.4834	0.877	0.168	0.056	0.429
Real-CamI2V [19]	1585.61	0.513	<u>0.3703</u>	<u>0.923</u>	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	<u>0.193</u>	0.068	<u>0.508</u>
Captain Safari	1023.46	0.512	0.3690	0.968	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 далее 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%	3,80%	4,47%
Captain Safari без памяти	25,00%	24,20%	24,20%	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 инлайнеров) для выявления потенциально недостоверных переходов. Затем *геометрическая проверка* пересматривает подозрительные пары, используя сохраненные ключевые точки и совпадения, пересчитывает основные матрицы и устанавливает пороги симметрических 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 поиском памяти, зависящим от позы, используя токены памяти, выровненные по позе m_t . Затем мы совместно обучаем поиском 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 остаётся 11,481 окончательных обучающих клипов. Дополнительно мы создаём 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.

Конфигурация и обозначения. Мы генерируем клипы $\mathcal{T} = 5$ 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 на шей базовой DiT с скрытым размером $D = 3072$. Извлекатель и DiT обучаются на 1 и 5 эпохах соответственно. Для каждого вideo мы извлекаем 3D-осведомленную характеристику памяти из предварительно обученной StreamVGTT [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.

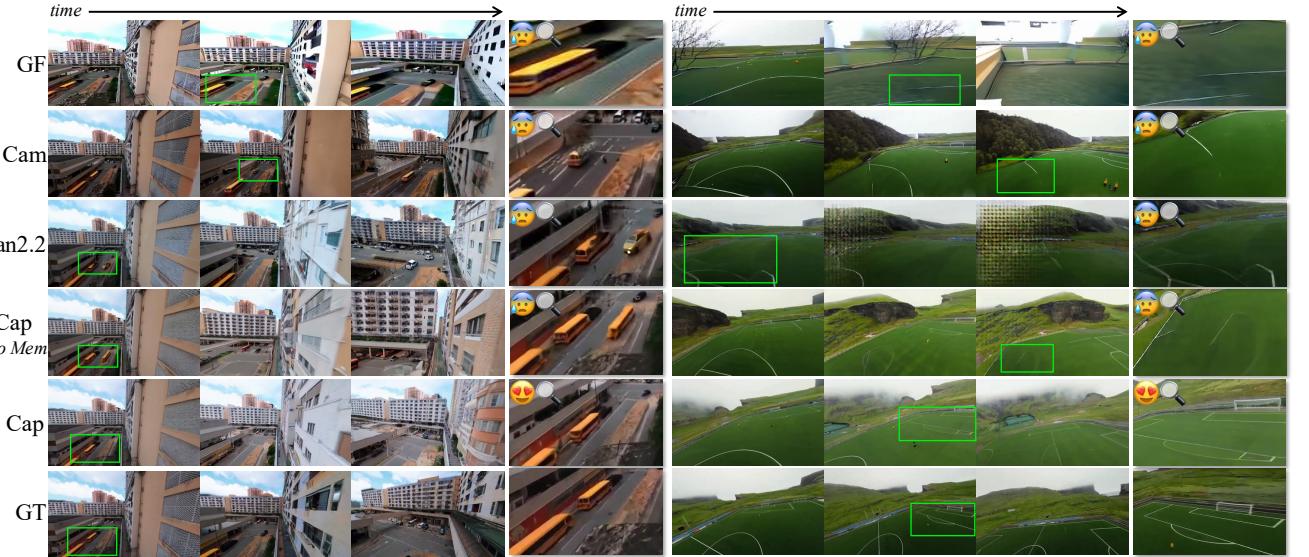


Figure 4. **Qualitative comparisons.** **Left:** Baselines—including the memory-removed variant—exhibit abrupt popping/vanishing of the school bus, and GF is low-quality. *Captain Safari* alone renders the bus smoothly exiting the frame. **Right:** Baselines distort or lose field marking, with Wan2.2 collapsing under large camera motion, affirming the challenge of 3D consistency under rapid trajectories. *Captain Safari* preserves crisp markings and coherent layout while following the fast 6-DoF path.

5.2. Benchmark

Metrics. We evaluate video generation along three complementary axes: video quality, 3D consistency, and trajectory following. For video quality, we report FVD [37] and LPIPS [54]. For 3D consistency, we use MEt3R [1], computed between GT and generated videos at matched time steps and a reconstruction rate that measures the percentage of frames successfully registered in the recovered 3D model [30, 31]. For trajectory following, we report camera relocation accuracy (AUC [39]) and the cosine similarity between the flattened camera pose, capturing how the model adheres to the desired camera parameters over time.

Baselines. We compare against representative camera-controllable video generation models, including Geometry Forcing [44], Real-CamI2V [19, 57], and Wan2.2-5B-Control-Camera [38], which cover geometry-constrained, reconstruction-driven, and large-scale diffusion-based approaches to trajectory-conditioned video synthesis.

Human Study. We conduct a human study with 50 participants. Each participant is presented with 10 cases, where each case contains the GT video and five anonymized model-generated videos (three baselines, our model, and its ablated variant). For every case, participants are asked to select the best video under three criteria: Video Quality, 3D Consistency, and Trajectory Following. In total, the study collects $50 \times 10 \times 3 = 1,500$ human preference votes.

5.3. Generation Quality

As shown in Table 1, our *Captain Safari* attains a substantially lower FVD (1023.46 vs. 1387.75) and a slightly

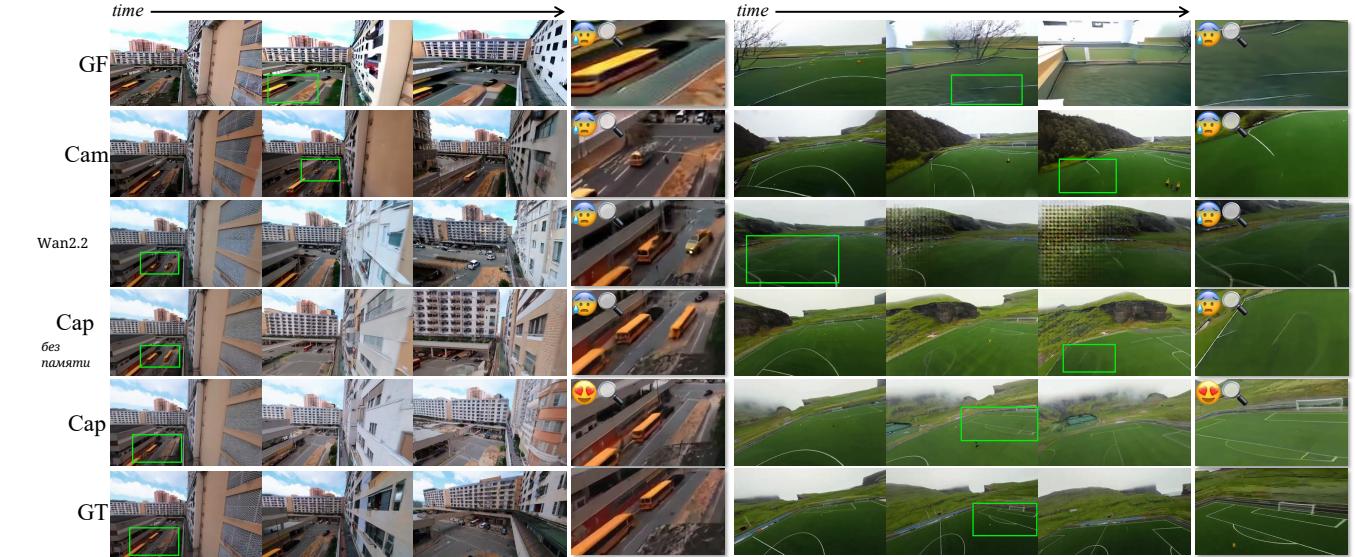


Рисунок 4. Качественные сравнения. Слева: Базовые модели — включая вариант без памяти — демонстрируют резкое появление/исчезновение школьного автобуса, а GF низкого качества. Только *Captain Safari* плавно выводит автобус из кадра. Справа: Базовые модели искажают или теряют разметку поля, при этом Wan2.2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

5.2. Бенчмарк

Метрики. Мы оцениваем генерацию видео по трем дополнительным осм: 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 реконструкции, 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 Geometry 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 *Captain Safari* достигает значительно более низкого FVD (1023.46 против 1387.75) и немногого

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 конкурентирующими методами, что подчеркивает перспективный 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 конкурентующие 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.

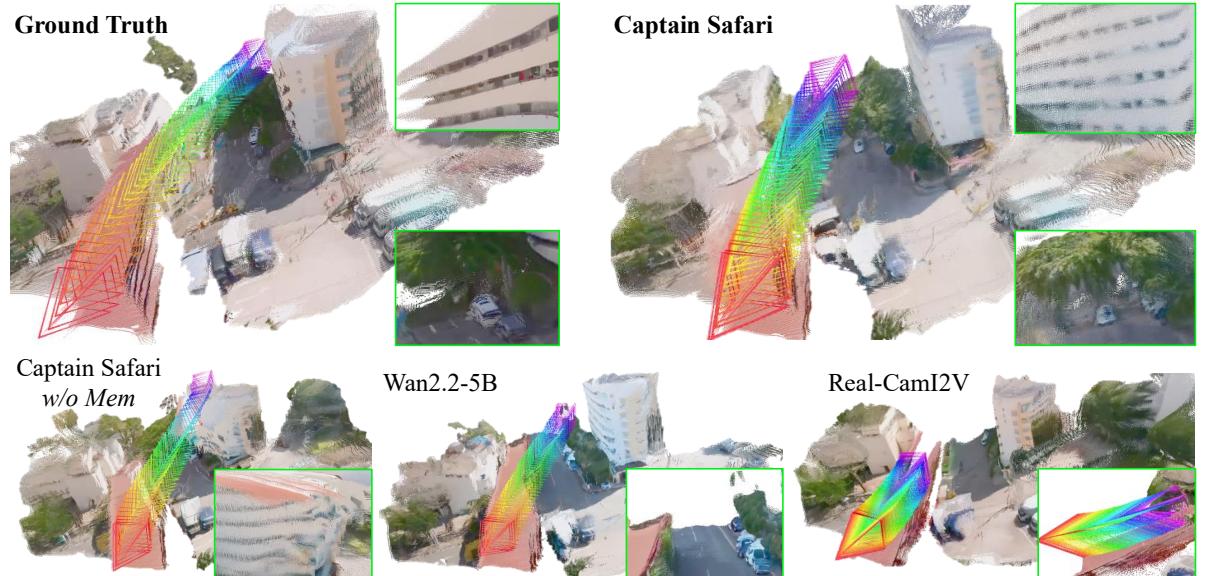


Figure 5. Scene reconstruction and camera trajectory. With pose-aligned memory, *Captain Safari* reconstructs a well-structured building façade (the memory-removed variant blurs/warps it), demonstrating the benefit of memory. It also preserves fine details—parked cars and the tree on their roofs—that Wan2.2-5B fails to retain. Meanwhile, Real-CamI2V follows only a short path, whereas *Captain Safari* covers the full trajectory with stable 3D structure, highlighting the challenge of maintaining 3D consistency under fast motion.

without collapsing geometry. Together, these results validate that the implicit world memory and pose-conditioned retrieval of *Captain Safari* effectively stabilize the underlying 3D world under aggressive camera motion.

5.5. Trajectory Following

Captain Safari delivers the most accurate trajectory following among all competing models. As shown in Table 1, our method achieves the highest AUC@30 (0.200) and AUC@15 (0.068), along with the best cosine similarity (0.563), outperforming the strongest baseline by clear margins. The human study in Table 2 further reinforces this observation, with **69.00%** of participants identifying our model as the most faithful to the target camera path.

Figure 5 provides a clear visualization of these improvements. *Captain Safari*'s predicted trajectory closely aligns with the ground-truth path, while the ablated variant deviates and flies over the rooftop, and RealCam-I2V fails to follow the intended forward motion, advancing only slightly rather than committing to the prescribed trajectory. Furthermore, our method demonstrates stable and coherent generation under challenging viewpoint changes with complex camera maneuvers in Figure 1. These results highlight the effectiveness of our memory-augmented, pose-conditioned design for precise trajectory adherence.

5.6. Ablation Study

Our results highlight the importance of the proposed pose-conditioned world memory. As shown in Table 1, adding memory yields substantial improvements in both 3D con-

sistency and trajectory following. These gains confirm that retrieving pose-aligned world features at the target frame provides the model with an explicit understanding of what the scene *should* look like, enabling stable geometry and accurate motion alignment.

Qualitative comparisons in Figure 4 and Figure 5 further illustrate these effects. With memory, the generated scenes preserve global structure, maintain consistent geometry across viewpoints. In contrast, the ablated variant often drifts and exhibits geometric inconsistencies. Together, these results validate the effectiveness of our memory-augmented design in stabilizing the underlying 3D world and guiding precise camera motion.

6. Conclusion

We introduced *Captain Safari*, a pose-conditioned world engine built on a world memory that enables long-range, 3D-consistent video generation under complex FPV trajectories. Together with *OpenSafari*, our curated dataset of in-the-wild drone videos with verified camera poses, this establishes a rigorous benchmark for controllable video generation. *Captain Safari* markedly improves 3D consistency and trajectory accuracy over prior methods while maintaining strong visual fidelity. Although the system incurs non-trivial inference overhead, future work will explore real-time world engines with lightweight memory and faster generative backbones. We hope *Captain Safari* and *OpenSafari* encourage further research in persistent world models and long-horizon controllable video generation.

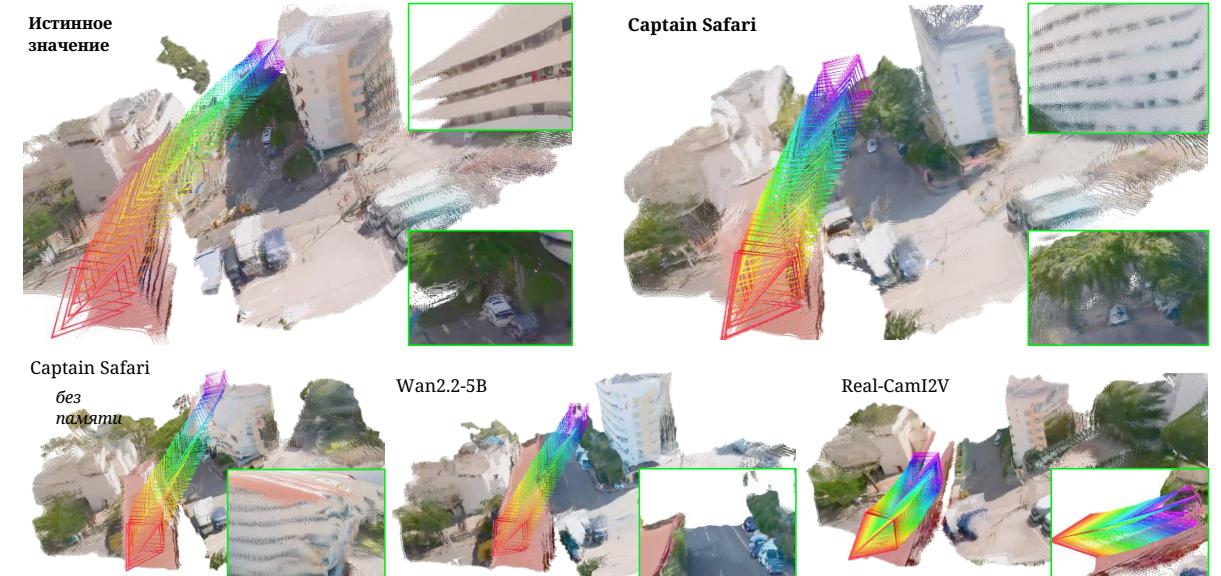


Рисунок 5. Реконструкция сцены и траектория камеры. С использованием памяти, выровненной по позе, *Captain Safari* реконструирует хорошо структурированный фасад здания (вариант без памяти размывает/искажает его), демонстрируя преимущество памяти. Он также сохраняет мелкие детали — припаркованные автомобили и дерево на их крышах — которые Wan2.2-5B не удаётся 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 базовую линию с 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 удаётся следовать заданному движению вперед, продвигаясь лишь немногим, 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 в стабилизации базового 3D мира и управлении точным движением камеры.

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 в

следовании траектории. Эти достижения подтверждают, что извлечение мировых признаков, выровненных по позе, в целевой кадр предоставляет модели явное понимание того, как сценарий 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. *Captain Safari* 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.

References

- [1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6034–6044, 2025. 7
- [2] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2024. 3
- [3] Shuai Bai, Kedqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [4] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoeffet, Brandon Houghton, Raul Sampredo, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *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, et al. Baking gaussian splatting into diffusion denoiser for fast and scalable single-stage image-to-3d generation and reconstruction. *arXiv preprint 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, et al. Visdrone-det2021: The vision meets drone object detection challenge results. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 2847–2854, 2021. 2
- [7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1
- [8] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. *arXiv preprint 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, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1–12, 2025. 3
- [10] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 1, 3
- [11] Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3
- [12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2
- Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6
- [13] Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheets: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12528–12537, 2021. 1, 2
- [14] Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhiuoao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint 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, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 3
- [16] Longbin Ji, Lei Zhong, Pengfei Wei, and Changjian Li. Pose-traj: Pose-aware trajectory control in video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22776–22785, 2025. 1, 2, 3
- [17] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024. 3
- [18] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024. 2
- [19] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yefan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28785–28796, 2025. 1, 3, 6, 7
- [20] Hanwen Liang, Junli Cao, Vedit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 798–810, 2025. 1, 3
- [21] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2
- [22] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European Conference on Computer Vision*, pages 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, et al. Genex: Generating an explorable world. *arXiv preprint arXiv:2412.09624*, 2024. 1, 2

Ссылки

- [1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Измерение многовидовой согласованности в генерированных изображениях. В материалах Международной конференции IEEE/CVF по компьютерному зрению, страницы 6034–6044, 2025. 7
- [2] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, 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, Kedqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl технический отчет. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [4] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoeffet, Brandon Houghton, Raul Sampredo 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, et al. Встраивание гауссового размытия в диффузионный денойзер для быстрой и масштабируемой однотапочной генерации и реконструкции изображений в 3d. *arXiv preprint 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 preprint arXiv:1709.06158*, 2017. 1
- [8] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang и Chuang Gan. Adaworld: Обучение адаптируемых мировым моделям с латентными действиями. *arXiv preprint 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 конференций по компьютерному зрению и распознаванию образов, страницы 1–12, 2025. 3
- [10] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yefan Xiong, Min Chen, et al. RealCam-i2V: Генерация видео из текста. *arXiv preprint arXiv:2404.02101*, 2024. 1, 3
- [11] Chen Hou и Zhibo Chen. Управление камерой для генерации видео без обучения. *arXiv preprint 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. CameraCtrl: Обеспечение управления камерой для генерации видео из текста. *arXiv preprint arXiv:2404.02101*, 2024. 1, 3
- [13] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2
- [14] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European Conference on Computer Vision*, pages 329–345. Springer, 2024. 2, 3
- [15] Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel Khashabi, Rama Chellappa, Alan Yuille и др. Genex: Generating an explorable world. *arXiv preprint arXiv:2412.09624*, 2024. 1, 2
- [16] Lora: Низкоранговая адаптация крупных языковых моделей. *ICLR*, 1(2):3, 2022. 6
- [17] Ronghang Hu, Nikhila Ravi, Alexander C Berg и Deepak Pathak. Worldsheets: 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
- [18] Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhiuoao Tian, Tianyu He, and Li Jiang. Memory Forcing: Пространственно-временная память для согласованной генерации сцен в Minecraft. *arXiv preprint arXiv:2510.03198*, 2025. 2, 3
- [19] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoeffet, Brandon Houghton, Raul Sampredo и Jeff Clune. 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
- [20] 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 preprint arXiv:2506.04225*, 2025. 3
- [21] 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
- [22] 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
- [23] 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
- [24] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yefan Xiong, Min Chen, et al. 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
- [25] Hanwen Liang, Junli Cao, Vedit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov 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
- [26] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov 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
- [27] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and 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
- [28] 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 preprint arXiv:2412.09624*, 2024. 1, 2

- [47] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3
- [48] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jia Shi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 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, et al. Omnicam: Unified multi-modal video generation via camera control. *arXiv preprint arXiv:2504.02312*, 2025. 2
- [50] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 1, 2
- [51] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5916–5926, 2025. 1, 3
- [52] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 3
- [53] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3
- [54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7
- [55] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025. 3
- [56] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. Motionpro: A precise motion controller for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27957–27967, 2025. 3
- [57] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 1, 3, 7
- [58] Mengqi Zhou, Yuxi Wang, Jun Hou, Shouga Zhang, Yawei Li, Chuanchen Luo, Junran Peng, and Zhaoxiang Zhang. Scenex: Procedural controllable large-scale scene generation. *arXiv preprint arXiv:2403.15698*, 2024. 2
- [59] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2, 5
- [60] Yunsong Zhou, Michael Simon, Zhenghao Peng, Sicheng Mo, Hongzi Zhu, Minyi Guo, and Bolei Zhou. Simgen: Simulator-conditioned driving scene generation. *Advances in Neural Information Processing Systems*, 37:48838–48874, 2024. 2
- [61] Zhenghong Zhou, Jie An, and Jiebo Luo. Latent-reframe: Enabling camera control for video diffusion models without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12779–12789, 2025. 3
- [62] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 6
- [63] 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
- [64] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jia Shi Feng и Hengshuang Zhao. Depth anything v2. Достижения в области систем обработки нейронной информации, 37:21875–21911, 2024. 3
- [65] 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
- [66] 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
- [67] 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
- [68] 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
- [69] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng 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
- [70] 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
- [71] 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
- [72] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao и Tao Mei. Motionpro: Точный контроллер движения для генерации изображений в видео. В материалах конференции по компьютерному зрению и распознаванию образов, страницы 27957–27967, 2025. 3
- [73] 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
- [74] Mengqi Zhou, Yuxi Wang, Jun Hou, Shouga Zhang, Yawei Li, Chuanchen Luo, Junran Peng и Zhaoxiang Zhang. Scenex: Процедурная управляемая генерация крупномасштабных сцен. *arXiv preprint arXiv:2403.15698*, 2024. 2
- [75] 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