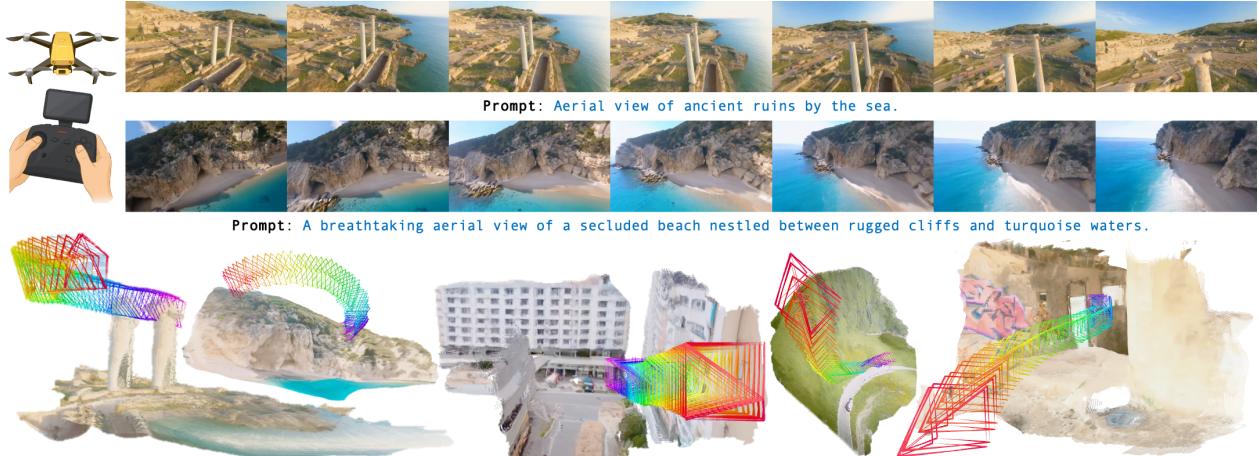


Captain Safari: Een Wereldmotor

Yu-Cheng Chou¹ Xingrui Wang¹ Yitong Li² Jiahao Wang¹ Hanting Liu¹ Cihang Xie³ Alan Yuille¹ Junfei Xiao¹[✉] Johns Hopkins Universiteit² Tsinghua University³
UCSantaCruz <https://johnson111788.github.io/open-safari/>



Figuur 1. CaptainSafari is een pose-bewuste wereldmotor die lange-horizon, 3D-consistente FPV-video's genereert vanuit elke door de gebruiker gespecificeerde cameratraject. Door pose-uitgelijnd wereldgeheugen op te halen, blijft de geometrie stabiel bij grote veranderingen in het gezichtspunt en worden scherpe, goed gevormde structuren gereconstrueerd terwijl agressieve 6-Dof bewegingen nauwkeurig worden gevolgd.

Samenvatting

Wereldmotoren zijn gericht op het synthetiseren van lange, 3D-consistente video's die interactieve verkennings van een scène ondersteunen onder door de gebruiker gecontroleerde camerabewegingen. Bestaande systemen hebben echter moeite met agressieve 6-Dof trajecten en complexe buitenlayouts: ze verliezen langeafstandsgeometrische samenhang, wijken af van het doelpad of vallen terug in te conservatieve bewegingen. Om deze reden introduceren we Captain Safari, een pose-geconditioneerde wereldmotor die video's genereert door te putten uit een persistent wereldgeheugen. Gegeven een camera-traject, behoudt onze methode een dynamisch lokaal geheugen en gebruikt een retriever om pose-uitgelijnde wereldtokens op te halen, die vervolgens de videogeneratie langs het traject conditioneren. Dit ontwerp stelt het model in staat om een stabiele 3D-structuur te behouden terwijl uitdagende cameramanoeuvres nauwkeurig worden uitgevoerd.

Om deze instelling te evalueren, hebben we OpenSafari samengesteld, een nieuw FPV-dataset in het wild met hoog-dynamische dronevideo's en geverifieerde cameratrajecten, geconstrueerd door middel van een meertraps geometrische en kinematische validatie.

[✉] Correspondentieauteur: Junfei Xiao (xiaojf97@gmail.com)
Preprint, werk in uitvoering.

lijn. Op het gebied van videokwaliteit, 3D consistentie en trajectvolg presteert CaptainSafari aanzienlijk beter dan de meest geavanceerde camera-gestuurde generatoren. Het vermindert MET3R van 0,3703 naar 0,3690, verbetert AUC@30 van 0,181 naar 0,200 en levert aanzienlijk lagere FVD op dan alle camera-gestuurde baselines. Belangrijker nog, in een menselijke studie met 50 deelnemers, waarbij annotatoren het beste resultaat kiezen uit vijf geanonimiseerde modellen, geeft 67,6% van de voorkeuren de voorkeur aan onze methode over alle assen. Onze resultaten tonen aan dat pose-geconditioneerde wereldgeheugen een krachtig mechanisme is voor lange-termijn, controleerbare videogeneratie en bieden OpenSafari als een uitdagende nieuwe benchmark voor toekomstig wereld-engine onderzoek.

1. Inleiding

Het simuleren van coherente 3D-werelden door middel van controleerbare videogeneratie is al lang een fundamentele uitdaging voor augmented reality, belichaamde AI en virtuele agenten [8–10, 13–16, 19, 20, 23, 26, 29, 40, 43–45, 50, 51, 57]. Klassieke game-engines en fysica-simulatoren bieden expliciete geometrie en nauwkeurige controle, maar vereisen veel handmatige bewerking en dure berekeningen [7, 28, 32]. Daar-

naast, hoewel ze visueel realisme kunnen bereiken in gespecialiseerde domeinen, schieten ze nog steeds tekort in het vastleggen van de rijkdom en diversiteit die kenmerkend zijn voor de echte wereld, zoals natuurlijke scènes [27, 35, 58]. In tegenstelling hiermee synthetiseren moderne videodiffusiemodellen hoogwaardige, diverse video's van tekst of afbeeldingen, maar functioneren ze meestal als feed-forward clipgeneratoren zonder blijvende wereldstatus: *ze hebben moeite met langeafstands 3D-consistentie, complexe trajectvolg en getrouwe reconstructie van diverse scènes* [18, 24]. In dit werk streven we ernaar deze kloof te overbruggen met *Captain Safari*, een wereldengine die pose-geconditioneerde modellering van 3D-consistente en diverse omgevingen mogelijk maakt, en de beperkingen van traditionele game-engines overtreft op het gebied van algemeenheid, diversiteit en interactiviteit.

Hedendaagse videowereldmodellen staan voor drie onderling verbonden uitdagingen. Ten eerste is *lange-termijn consistentie* beperkt door het temporele venster van contextframes; modellen “vergeten” vaak verre landschappen of schenden ruimtelijke samenhang, wat leidt tot abrupte veranderingen in uiterlijk die de realisme en continuïteit van de gegenereerde omgeving doorbreken [8, 14, 45]. Ten tweede blijft het moeilijk om *complex camerabewegingen onder strikte 3D consistentie* te bereiken: bestaande methoden die afhankelijk zijn van pose- of trajectconditionering werken meestal goed alleen voor langzame, bijna voorwaartse bewegingen [16, 34, 49]. Wanneer het pad snelle 6-DoF-beweging, sterke parallax of scherpe bochten omvat, vertonen modellen een afweging—of wel het dempen van beweging en het beperken van veranderingen in het gezichtspunt om de geometrie te behouden, ofwel het volgen van het gevraagde pad ten koste van vervormingen, flikkeringen en structurele afwijkingen. Ten derde worden *complex buitenindelingen* momenteel ondervertegenwoordigd. Veel van het werk richt zich op gestructureerde, beperkte omgevingen (bijv. indoor tours, risituaties of vastgoedvideo's), en modellen worden zelden onderworpen aan stress-tests in wilde FPV-scenario's waar de camera zich tussen gebouwen, vegetatie en gevarieerd terrein met aanzienlijke parallax beweegt [6, 22, 59, 60]. Als gevolg daarvan falen methoden die competitief lijken in vereenvoudigde omgevingen vaak om geometrie en uiterlijk te behouden wanneer ze worden geconfronteerd met echt diverse, complexe buitenomgevingen.

Om deze problemen aan te pakken, introduceren we *Captain Safari*, een pose-bewuste wereldmotor die expliciet een blijvend begrip van de wereldtoestand handhaaft om *langetermijn 3D consistentie* te behouden bij sterke parallax. Omdat het opslaan en doorgeven van een volledige langetermijnstaat computationeel onhaalbaar is, ontwikkelen we een ophaalmechanisme dat *alleen de meest informatieve scène-aanwijzingen selecteert en aggregateert*, waardoor sterke geometrische begeleiding wordt geboden zonder onaanvaardbare kosten. Cruciaal is dat deze ophaalactie *pose-bewust* is: gegeven de doelcamerapositie, stelt het een pose-uitgelijnde wereldvoorafbeelding samen die het generatieproces stuurt, waardoor nauwkeurige tracking van *agressieve camerabewegingen* mogelijk wordt gemaakt, terwijl 3D-consistente structuren in complexe omgevingen behouden blijven.

Bovendien, om de kloof te dichten in *complex buitenomgevingen* en *agressieve camerabewegingen*, hebben we *OpenSafari* samengesteld,

een grootschalig dataset van hoog-dynamische FPV dronevideo's met geverifieerde cameraposities. Veel van de literatuur richt zich op gestructureerde, beperkte omgevingen (bijv. indoor tours, rij- of vastgoedvideo's), en zelfs buitendatasets bevatten meestal langzame, bijna voorwaartse bewegingen. Daarentegen omvat *OpenSafari* FPV-vluchten in de vrije natuur die zich om gebouwen en vegetatie heen bewegen over oneffen terrein, met grote parallax, snelle 6-DoF manoeuvres en scherpe veranderingen in gezichtspunt. In combinatie met geverifieerde cameratrajecten presenteren deze video's diverse, rommelige buitenscènes en langeafstandsbewegingen, wat modellen uitdaagt om 3D-consistentie te behouden terwijl ze complexe manoeuvres nauwkeurig volgen.

We evalueren *Captain Safari* langs drie assen: *videokwaliteit, 3D consistentie, entrajectvolg*. Volgens deze criteria presteert onze methode consequent beter dan hedendaagse cameragestuurde videogeneratoren op *OpenSafari*: Tabel 1 rapporteert duidelijke verbeteringen in 3D consistentie en nauwkeurige tracking onder complexe manoeuvres, terwijl sterke perceptuele kwaliteit behouden blijft. Belangrijk is dat een grootschalige menselijke studie (Tabel 2) aantoont dat *Captain Safari* **67% van de stemmen ontvangt in vergelijkingen met vijf opties, wat aangeeft dat de verbetering perceptueel opvallend zijn**. **Kwalitatieve vergelijkingen** (Fig. 4 en Fig. 5) tonen verder stabiele geometrie aan bij langeafstandspaden en getrouwe naleving van scherpe 6-DoF camerabewegingen in rommelige buitenomgevingen.

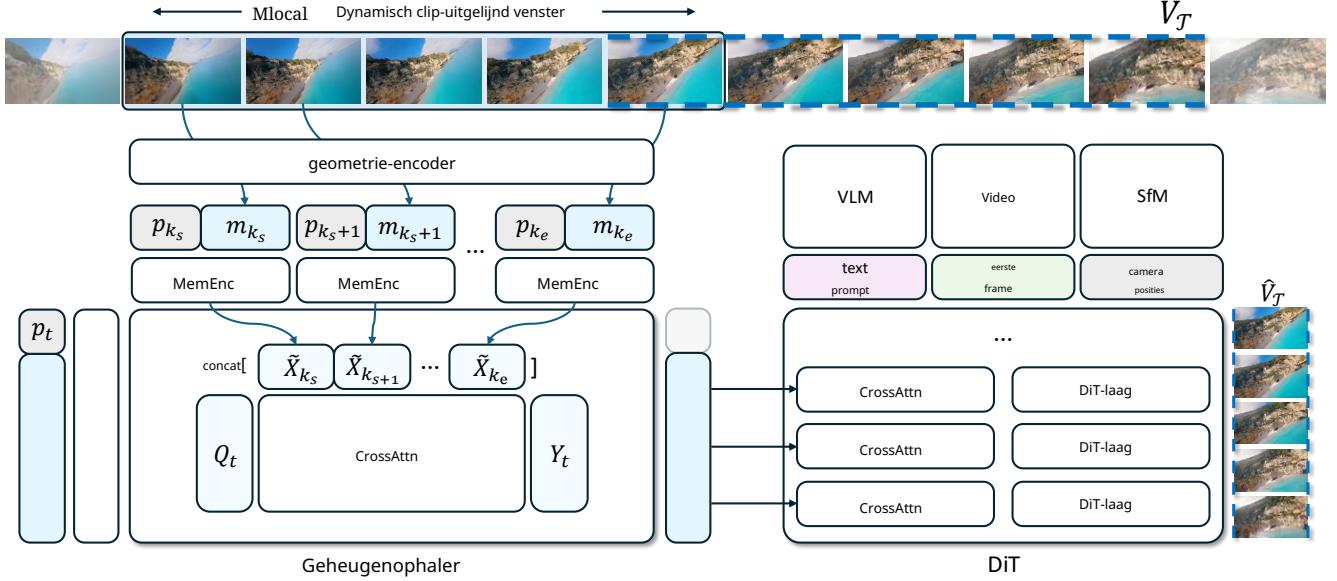
Samenvattend zijn onze bijdragen:

1. We presenteren *Captain Safari*, de eerste cameragestuurde videogeneratiemethode die lange-termijn 3D consistentie afdwingt terwijl agressieve FPV-bewegingen worden gevuld. 2. We stellen een *pose-geleide, lange-termijn retrieval* voor die efficiënt strikte 3D consistentie verzoent met nauwkeurige tracking van complexe manoeuvres. 3. We stellen *OpenSafari* samen, een grootschalige in-the-wild FPV-dataset met geverifieerde cameraposities, met diverse, rommelige buitenscènes en snelle 6-DoF bewegingen die de geometrie-consistente camerabesturing op de proef stellen. 4. In *OpenSafari* verbetert onze pose-bewuste retrieval opmerkelijk de videokwaliteit, 3D consistentie en trajectuitlijning, en behaalt ook een **67%** menselijke voorkeursscore.

2. Gerelateerd Werk

2.1. 3D-consistente wereldmodellen

Vroege benaderingen voor beeld-naar-3D reconstrueren geometrie indirect via multi-view consistentie of impliciete velden, maar slagen er vaak niet in om een coherente structuur te behouden bij grote veranderingen in het zicht [13, 21, 43, 50]. Recente inspanningen integreren 3D-redenering in het generatieve proces. DiffusionGS [5] injecteert Gaussian S plattening in de diffusie denoiser, waardoor zichtconsistentie wordt afgedwongen en schaalbare 3D-generatie in één stap mogelijk wordt. GenEx [23] en EvoWorld [40] breiden dit idee uit van statische reconstructie naar dynamische wereldcreatie, waarbij verkenbare 360° panoramische omgevingen worden gegenereerd.



Figuur 2. **Methodeoverzicht.** *Captain Safari* bouwt een lokaal wereldgeheugen op en haalt, gegeven een query camerapositie, pose-uitgelijnde tokense op die de scène samenvatten. Deze tokense conditioneren vervolgens de videogeneratie langs het door de gebruiker gespecificeerde traject, waarbij een stabiele 3D-indeling behouden blijft.

gebaseerd op fysieke aannames. Complementair aan deze generatieve reconstructies, Geometrie Dwang [44] en Geheugenforcing [14] koppelen trainingssignalen expliciet met geometrische supervisie en spatio-temporeel geheugen, wat consistentie tijdens lange uitrol garandeert. Ondertussen integreren open-wereldmodellen zoals Wonderland [20], Wonderland [51], Wonder-Turbo [26], en EvoWorld [40] verder geometrie-geïndexeerde of adaptieve geheugens om blijvende wereldstaten te behouden tijdens interacties. Deze benaderingen gebruiken echter nog steeds impliciete, clip-gebonden geheugens, terwijl wij een expliciet pose-geïndexeerd wereldgeheugen introduceren dat op aanvraag wordt opgevraagd voor camera-gestuurde generatie.

2.2. Cameragestuurd Videogeneratie

Vroege T2V/I2V-modellen leerden camerabeweging impliciet en hebben moeite om expliciete trajecten betrouwbaar te herhalen [11, 42, 61]. Recent werk zoals CameraCtrl [10] behandelt cameraparameters als expliciete voorwaarden, waarbij camera-extrinsieken en trajecten worden gecodeerd of padbeperkingen worden afgedwongen—om de bestuurbaarheid en nauwkeurigheid te verbeteren [2, 25, 41, 47, 55]. Motion-Prompting [9] implementeert compositiecontrole door punt-track conditionering en MotionPro [56] gebruikt pad-uitlijningsverliezen die de rotatie- en translatifout verminderen; controle zonder training wordt ook bereikt door een lichte puntenwolk aan te passen en een ruis-layout prior te gebruiken om het denoisen te sturen [11]. Scènebehoudende geometrische priors versterken verder de clip-niveau consistentie. Cami2V[57] behandelt camerapositie als een fysieke prior en benut epipolaire en multiview-beperkingen; RealCam-I2V [19] herstelt metrische diepte met DepthAnything v2 [48] om een schaalstabiele scène te reconstrueren; PoseTraj[16] maakt gebruik van pose-bewuste pretraining om

rotatie-uitgelijnde beweging te verkrijgen. vergeleken met alleen parameterconditionering verminderen deze priors de layout-drift binnen clips en behouden ze beter de lokale geometrie bij veranderingen in het zicht. Verder koppelt recent werk camerabesturing aan wereldmodellering, CVD [17], Cavia [46], en WoVoGen [22] syntheseren gezamenlijk multi-view en multi-traject video's vanuit een gedeelde scène-representatie, waarbij cross-pad consistentie wordt afgedwongen. Ondertussen kunnen methoden die conditioneren vanuit expliciete renderbare 3D-representaties (bijv. 3D Gaussians) geometrie verankeren, cross-view 3D consistentie verbeteren en padnauwkeurigheid bevorderen [15, 29, 33, 52, 53]. Deze benaderingen bouwen echter meestal eenmalige 3D-scènes, terwijl wij lange-termijn camerabesturing verenigen met een persistent pose-geïndexeerd wereldgeheugen dat wordt gedeeld over trajecten.

3. Captain Safari

We introduceren *Captain Safari*, een geheugen-gestuurde videogeneratie framework. Sectie 3.1 presenteert een impliciet wereldgeheugen voor stabiele scenerepresentatie, terwijl Sectie 3.2 een pose-geconditioneerd ophaalsysteem beschrijft dat camerabeelden naar wereldtokense mapt, wat een DiT-gebaseerde generator leidt voor coherente outputs langs willekeurige trajecten.

3.1. Impliciet Geheugen van Wereldgeometrie

Probleemopstelling. We representeren een video als $V = \{I_t\}_{t=0}^T$, waarbij I_t het frame is op tijdstip t . Op dezelfde tijdsas definiëren we cameraposities $\mathcal{C} = \{(R_t, T_t)\}_{t=0}^T$ en verkrijgen we een 3D-bewust geheugenkenmerk m_t op elke tijdstip t met behulp van een voorgetrainde geometrie-encoder. Alle geheugenkenmerken vormen een globale geheugenbank $\mathcal{M} = \{m_t\}_{t=0}^T$.

Gegeven een teksprompt p , de cameraposities \mathcal{C} , en een doel

clip tijdstap $\mathcal{T} = [t_0, t_1]$, samen met het bijbehorende lokale wereldgeheugen $\mathcal{M}_{\text{local}} \subset \mathcal{M}$, is ons doel om een videosegment $V_{\mathcal{T}}$ te synthetiseren dat (i) overeenkomt met p , (ii) de voorgeschreven posities respecteert $\{(R_t, T_t)\}_{t \in \mathcal{T}}$, en (iii) een samenhangende 3D-wereld behoudt over verschillende gezichtspunten.

Lokale wereldgeheugen. Direct conditioneren op de volledige geheugenbank \mathcal{M} voor elke clip zou computationeel duur zijn en gedomineerd worden door temporele verre observaties. In plaats daarvan definiëren we voor elke doel-clip tijdstap $\mathcal{T} = [t_0, t_1]$ een *lokaal* geheugen $\mathcal{M}_{\text{local}} = \{m_{\tau} \mid \tau \in [k_s, k_e]\}$ waarvan de eindpunten worden bemonsterd onder

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

waar L een vaste grens is en alle tijdstappen gehele getallen zijn. Deze beperkingen zorgen ervoor dat: (i) het geheugenvenster begint maximaal L seconden voor de clipingang t_0 , waardoor het aan nabije observaties wordt gekoppeld; (ii) de duur is maximaal L , wat de conditionerende set compact houdt; en (iii) de eindtijd k_e alt altijd samen met of overlapt t_0 terwijl het binnen $[t_0, t_1]$ blijft, wat ervoor zorgt dat elke clip wordt ondersteund door een temporeel compatibele wereldvooraaf. Alle $\mathcal{M}_{\text{local}}$ worden geconstrueerd als een dergelijk dynamisch clip-uitgelijnd venster van de gedeelde bank \mathcal{M} , zodat naburige clips van nature overlappende geheugenvermeldingen delen, wat de berekening beperkt terwijl hun generaties worden gekoppeld aan een 3D-consistente onderliggende wereld.

Pose-opgehaald geheugen. Binnen een gegeven clip tijdstap \mathcal{T} beschouwen we het lokaal geheugen $\mathcal{M}_{\text{local}}$ als een statische hypothese van de omringende wereld, opgebouwd uit sleutelbeelden. Elke tijdstap τ levert een pose-token p_{τ} (afgeleid van (R_{τ}, T_{τ})) en een set van 3D-bewuste geheugentokens $m_{\tau,1}, \dots, m_{\tau,M}$. De verzameling $\{(p_{\tau}, m_{\tau,1:M})\}_{\tau}$ vormt een impliciete werelddatabase: het pose-token geeft aan waar de camera de scène heeft waargenomen, terwijl geheugentokens coderen hoe de wereld eruitziet vanuit die configuraties. Voor elke doel-tijdstap $t \in \mathcal{T}$ leiden we de camerapositie af naar een query pose-token p_t , embedden het als $q_t = \phi_p(p_t)$, en gebruiken een speciale ophaalmodule om op een pose-afhankelijke manier uit deze statische tabel te lezen. Concreet wordt q_t geconcateneerd met een bank van leerbare querytokens en verwerkt tot ophaalqueries, die cross-attentie uitvoeren over het gecodeerde geheugen X^{mem} (gedefinieerd in Sectie 3.2), wat resulteert in een set van wereldtokens.

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

overeenkomend met de bijgewerkte leerbare queries. Deze pose-uitgelijnde wereldtokens w_t worden direct gebruikt als het gereconstrueerde geheugen op pose t . Zo krijgen alle frames in \mathcal{T} toegang tot lokaal geheugen via pose-geconditioneerde queries in plaats van ruwe tijdsindices, wat aanmoedigt dat multi-view observaties verbonden blijven met een consistent statische 3D-wereld.

3.2. Geheugenophaling en Conditioning

Geheugenophaalontwerp. Zoals getoond in Figuur 2, gegeven het lokaal geheugen, representeren we elke tijdstap τ door een pose

token p_{τ} en de bijbehorende geheugentokens $m_{\tau,1:M}$. Onze ophaler is ontworpen om (i) pose-geheugentokens gezamenlijk te coderen in een samenhangende wereldrepresentatie, en (ii) voor elke querypose een compacte set pose-uitgelijnde tokens te extraheren die de meest relevante delen van deze lokale wereld samenvatten.

We embedden eerst pose- en geheugentokenmerken in een gedeelde ruimte en vormen een gezamenlijke reeks per tijdstap:

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

waarbij ϕ_p en ϕ_m respectievelijk leerbare embeddings voor pose- en geheugentokens aanduiden. Een stapel transformerblokken (MemEnc) met 3D-bewuste positionele codering verfijnt deze reeksen,

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

en we verkrijgen het gecodeerde lokaal wereldgeheugen door concatenatie

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

optioneel gemaskeerd om opgevulde of niet-sleutelitems uit te sluiten.

Voor een doel-tijdstap t leiden we de query pose-token p_t af, embedden deze als $q_t = \phi_p(p_t)$, en voegen deze samen met M leerbare query-tokens r_1, \dots, r_M ,

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

Deze reeks wordt verfijnd door transformerblokken die dezelfde architectuur delen als MemEnc, aangeduid als QryEnc, wat leidt tot pose-bewuste ophaalqueries

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

We voeren vervolgens cross-attentie uit van Q_t naar het gecodeerde geheugen X^{mem} ,

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

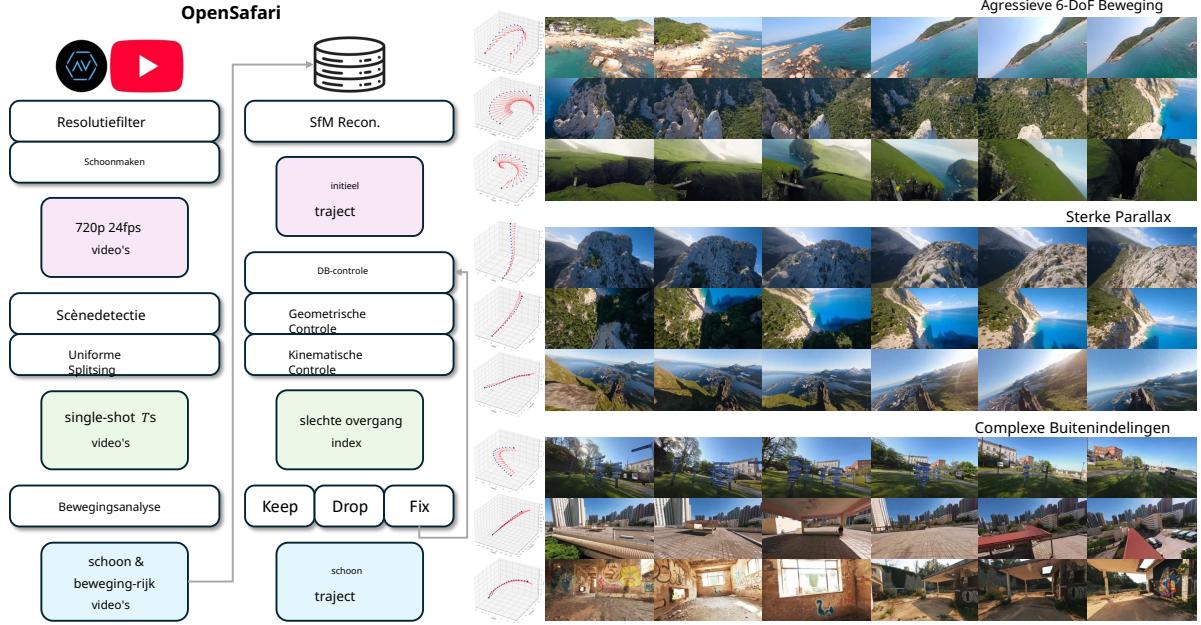
en nemen de subset van tokens in Y_t die overeenkomen met de leerbare queries als de opgehaalde wereldtokens

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

die een pose-uitgelijnde wereldfunctie vormen voor tijd t . Tijdens de training map een lineaire kop w_t terug naar de oorspronkelijke geheugenuimte om de doelgeheugentokens bij de query pose te reconstrueren. Het stappen van meerdere ophaalblokken verfijnt iteratief zowel de queries als de opgehaalde tokens, waardoor het model in staat is om elke query pose zachtjes te routeren naar de meest relevante subset van eerdere observaties, in plaats van te vertrouwen op een rigide temporele buurt of een enkel dichtstbijzijnd frame.

Geheugen-geconditioneerde DiT. Voor een gegeven doel-clip tijdstap \mathcal{T} , verwerkt de retriever $\mathcal{M}_{\text{local}}$ en de queryhouding p_t en levert een pose-uitgelijnde set van wereldtokens $w_t \in \mathbb{R}^{M \times d_m}$, die de statische lokale wereld samenvatten die relevant is voor dit segment. Deze tokens worden in de verborgen ruimte van DiT gemapt door de geheugen-embedding MLP.

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



Figuur 3. *OpenSafari*. Een nieuwe FPV-dataset in de natuur met zorgvuldig geverifieerde cameratrajecten, ontworpen om geometrie-consistente, camera-controleerbare videogeneratie te testen. We selecteren clips via een compacte, meervoudige pijplijn die trajecten filtert, reconstrueert en verifieert, resulterend in schone, bewegingrijke video's met betrouwbare camerapaden.

De latente clip wordt gecodeerd als een enkele spatio-temporele tokenreeks $Z \in \mathbb{R}^{L_z \times D}$, verkregen door alle frames in V_T te patchen. Bij elke DiT-laag l passen we eerst zelf-attentie toe over de volledige reeks en injecteren vervolgens de wereldtokens via een speciale geheugen-cross-attentie:

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

De wereldtokens op clipniveau W_T worden hergebruikt als sleutels en waarden over alle lagen, wat een stabiele, 3D-consistente basis biedt die de ruisonderdrukking van elke spatio-temporele token vormgeeft.

4. OpenSafari

4.1. Video Data Curation

Bestaande camera-geconditioneerde datasets komen niet overeen met ons doelregime. RealEstate10K [59] richt zich op langzame, meestal binnenuit-wandelingen met zachte bewegingen en schone, quasi-statistische scènes, terwijl Minecraft [4] een synthetische voxelwereld is met vereenvoudigde geometrie en door de engine beperkte dynamiek. Geen van beide legt agressieve, in-the-wild 6-DoF dronevluchten vast met sterke parallax, grote hoogteverschillen en complexe buitenindelingen die echt de lange-termijn 3D consistentie testen. Daarom stellen we *OpenSafari* voor, een nieuwe dataset van real-world FPV-stijl dronevideo's met geverifieerde cameratrajecten die zijn afgestemd op deze uitdagende omgeving.

We construeren Safari-FPV uit FPV-stijl dronevideo's

verzameld op AirVuz¹ en YouTube², en behouden alleen clips die een strikte meervoudige voorverwerkingspijplijn doorstaan. Zoals getoond in Figuur 3, doen we het volgende: (i) we downloaden de hoogste beschikbare resolutie voor elke URL en verwijderen bronnen onder de doelresolutie; (ii) we normaliseren alle video's naar 720p, 24fps, en een vaste 16:9 centrale uitsnede, waarbij we letterboxen en zwarte randen verwijderen zodat de daaropvolgende camera-estimatie op een schoon gezichtsveld werkt; (iii) we voeren scènedetectie uit om enkelvoudige segmenten te verkrijgen; (iv) we splitsen segmenten in video's van vaste lengte via uniforme temporele slicing.

We filteren vervolgens video's met een enkele diagnose op basis van beweging. Specifiek draaien we RAFT [36] om de optische-stroomgrootte te schatten; video's met te weinig beweging worden verwijderd, terwijl video's met stabiele, coherente beweging worden behouden om informatieve, parallax-rijke trajecten te benadrukken in plaats van statische beelden. Alleen video's die voldoen aan de bewegingsbeperking komen in de uiteindelijke dataset. Dit levert een grootschalige, in-the-wild dronecorpus op die expliciet is afgestemd om geometriebewuste, trajectvolgende videogeneratie te stress-testen.

4.2. Camera Trajectoireconstructie

Voor elke samengestelde video schatten we de interne en externe camerakenmerken bij 4fps met behulp van Hiërarchische Lokalisatie [30, 31]. We extraheren lokale kenmerken, bouwen uitputtende beeldparen binnen elke video, voeren kenmerkmatching uit en reconstrueren een C-OLMAP-stijl SfM-model; uit dit model exporteren we

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

Tabel 1. **Benchmark camera-gestuurde videogeneratie.** *Captain Safari* staat op de eerste plaats in 3D consistentie en trajectvolgning met concurrerende videokwaliteit. vergeleken met de geablateerde variant zonder geheugen, verbetert *Captain Safari* de 3D consistentie en trajectvolgning aanzienlijk, met slechts een kleine concessie in videokwaliteit. (Recon. = reconstructiesnelheid. CosSim = cosinusgelijkenis.)

Model	Videokwaliteit		3D consistentie		Trajectvolgning		
	FVD ↓	LPIPS ↓	MEt3R ↓	Recon. ↑	AUC@30 ↑	AUC@15 ↑	CosSim ↑
Geometrie Dwang [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 z/oGeheugen.	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

Tabel 2. **Menselijke voorkeur.** Gebruikers geven overweldigend de voorkeur aan *Captain Safari* op alle criteria, met 67% van de totale stemmen. De variant zonder geheugen staat op een verre tweede plaats, terwijl de basisconcurrenten een voorkeur in enkelcijferige percentages ontvangen.

Model	Videokwaliteit	3D Consistentie	Trajectvolgning	Gemiddeld
Geometrie Dwang [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 z/oGeheugen.	25,00%	24,20%	20,00%	23,07%
Captain Safari	67,40%	65,60%	69,00%	67,33%

per-frame cameragegevens als initiële trajecten.

Om gegevens gereed voor implementatie te verkrijgen, passen we een verificatie- en herstelpipeline in drie fasen toe op elke gereconstrueerde trajectorie. Eerst gebruikt *databasecontrole* SfM-statistieken (inlier-aantallen en -verhoudingen) om potentieel onbetrouwbare overgangen te markeren. Vervolgens herbekijkt *geometrische controle* verdachte paren met behulp van opgeslagen sleutelpunten en overeenkomsten, herberekent essentiële matrices en stelt drempels voor symmetrische epipolaire fouten. Ten slotte analyseert *kinematische controle* de posevolgorde op vertaalpijken, rotatiesprongen, omkeringen van de voorwaartse richting en schendingen van hogere-orde gladheid, waarbij robuuste MAD-gebaseerde scores worden gebruikt om onwaarschijnlijke bewegingen te detecteren.

De beslissingen per overgang worden samengevoegd tot een binaire slechte-index, die een strikt beleid aanstuurt. Als slechte overgangen schaars en lokaal zijn, passen we een gerichte correctie toe: we interpoleren lineair de cameracentra en passen SLERP toe op rotaties met een begrenste interpolatiehoek, waarbij we desgewenst extrapoleren aan de videoranden. De gecorrigeerde segmenten worden vervolgens opnieuw gevalideerd volgens dezelfde database-/geometrische-/kinematische criteria. Als de validatie na correctie slaagt, wordt de trajectorie geëxporteerd naar het definitieve dataset. Als de slechte-index te dicht is, de schendingen te ernstig zijn, of de gecorrigeerde trajectorieën nog steeds niet slagen voor verificatie, wordt de gehele video verworpen.

Het resulterende *OpenSafari* combineert dynamische, in-het-wild FPV drone video met rigoureus geverifieerde camera-trajecten. Het wijkt af van bestaande benchmarks door de nadruk te leggen op agressieve 6-DoF beweging, sterke parallax en complexe buitenlayouts, terwijl strikte geometrische en kinematische validatie wordt afgedwongen. Dit maakt *OpenSafari* een uitdagende testomgeving voor camera-controleerbare videoproduktie.

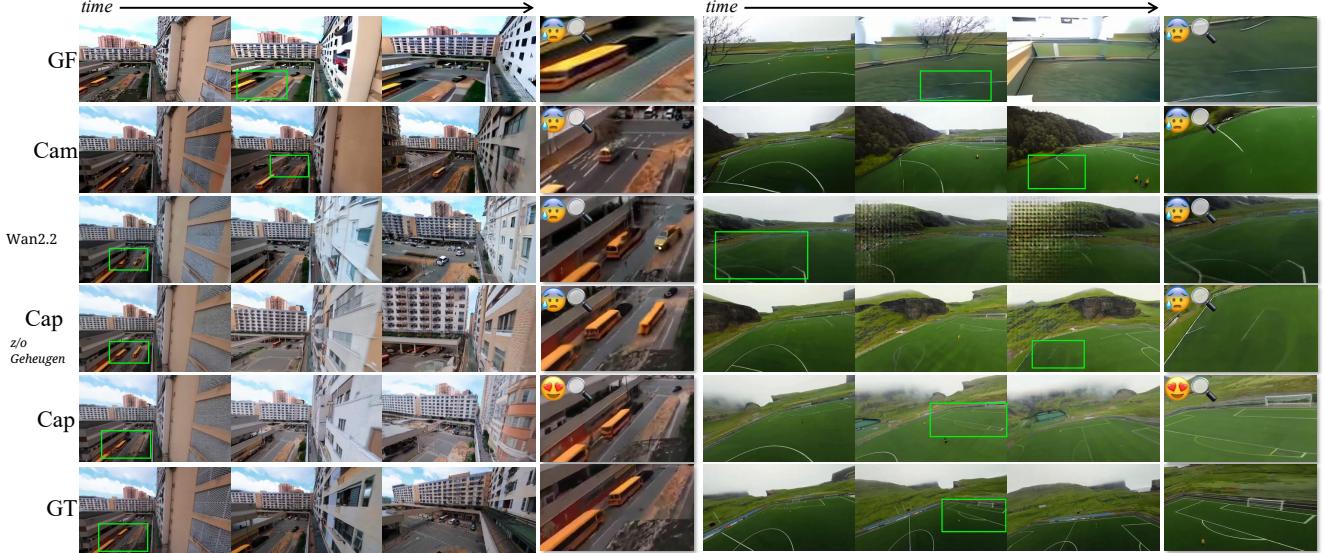
5. Experimenten

5.1. Implementatiедetails

Trainingsrecept. We hanteren een tweestapsrecept. Eerst warmen we de pose-geconditioneerde geheugenophaler op met behulp van pose-uitgelijnde geheugentokens m_t . Vervolgens trainen we de ophaler en DiT gezamenlijk end-to-end, waarbij we de DiT bijwerken via LoRA [12]. Geheugen cross-attentie wordt geïnitialiseerd vanuit de overeenkomstige context cross-attentie gewichten, en andere nieuwe lagen gebruiken standaardinitialisatie.

Dataset. We extraheren overlappende clips met een stap van 1s, resulterend in 51,997 trainingskandidaten. Een diversiteitsgebaseerd trajectfilter verwijdert clips met bijna-statische beweging, wat resulteert in 11,481 definitieve trainingsclips. We construeren daarnaast een niet-overlappende testset van 787 clips voor evaluatie. Voor elke clip genereren we een enkele beschrijvende caption met behulp van Qwen2.5-VL-7B [3] en gebruiken deze als tekstvoorwaarde.

Configuratie en notatie. We genereren $\mathcal{T} = 5$ s clips bij 24 fps van $T = 15$ s video's. Camerapositie en geheugenkenmerken worden gesampled bij 4 fps. Voor een doelclip van 5 s met interval $[t_0, t_1]$ gebruiken we de eindpositie p_{t_1} als de query. Het geheugenvenster is beperkt tot $L = 5$ s. We gebruiken Wan2.2-Fun-5B-Control-Camera [38] als onze basis DiT met een verborgen dimensie $D = 3072$. Retriever en DiT worden respectievelijk getraind met 1 en 5 epochs. Voor elke video extraheren we 3D-bewuste geheugenkenmerken van een voorgetrainede StreamVGTT [62]. We selecteren vier lagen $\{4, 11, 17, 23\}$; op elke laag bevattet het kenmerk 782 tokens. Het samenvoegen over de vier lagen levert $M = 4 \times 782$ en $d_m = 1024$ geheugentokens per frame op.



Figuur 4. **Kwalitatieve vergelijkingen.** Links: Baselines—waaronder de variant zonder geheugen—vertonen abrupt opduiken/verdwijnen van de schoolbus, en GF is van lage kwaliteit. Captain Safari alleen laat de bus soepel uit het frame verdwijnen. Rechts: Baselines vervormen of verliezen veldmarkeringen, waarbij Wan2.2 instort bij grote camerabewegingen, wat de uitdaging van 3D consistentie onder snelle trajecten bevestigt. Captain Safari behoudt scherpe markeringen en een samenhangende lay-out terwijl het het snelle 6-DoF pad volgt.

5.2. Benchmark

Metrieken. We evalueren videogeneratie langs drie complementaire assen: videokwaliteit, 3D consistentie en trajectvolging. Voor videokwaliteit rapporteren we FVD [37] en LPIPS [54]. Voor 3D consistentie gebruiken we MEt3R [1], berekend tussen GT en gegenereerde video's op overeenkomende tijdstappen en een reconstructiesnelheid die het percentage frames meet dat succesvol is geregistreerd in het herstelde 3D-model [30, 31]. Voor trajectvolging rapporteren we de nauwkeurigheid van cameraverplaatsing (AUC [39]) en de cosinusgelijkenis tussen de afgevlakte camerapositie, die vastlegt hoe het model zich houdt aan de gewenste cameraparameters in de loop van de tijd.

Baselines. We vergelijken met representatieve camera-controleerbare videogeneratiemodellen, waaronder Geome-try Forcing [44], Real-CamI2V [19, 57], en Wan2.2-5B-Control-Camera [38], die geometrie-gebonden, reconstructie-gedreven en grootschalige diffusie-gebaseerde benaderingen voor traject-geconditioneerde videosynthese omvatten.

Menselijke Studie. We voeren een menselijke studie uit met 50 deelnemers. Elke deelnemer krijgt 10 gevallen gepresenteerd, waarbij elk geval de GT-video en vijf geanonimiseerde door modellen gegenereerde video's bevat (drie baselines, ons model en de geablateerde variant). Voor elk geval wordt de deelnemers gevraagd de beste video te selecteren op basis van drie criteria: Videokwaliteit, 3D Consistentie en Trajectvolging. In totaal verzamelt de studie $50 \times 10 \times 3 = 1,500$ menselijke voorkeurstemmen.

5.3. Generatiekwaliteit

Zoals getoond in Tabel 1, behaalt onze CaptainSafari een aanzienlijk lagere FVD (1023,46 vs. 1387,75) en een iets

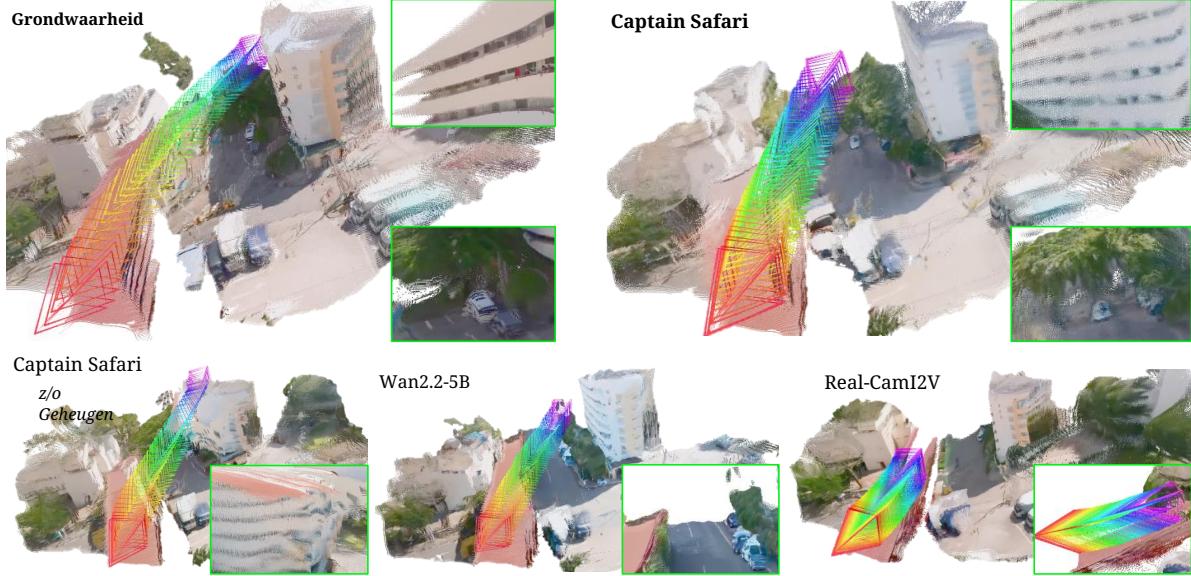
verbeterde LPIPS-score (0,512 vs. 0,513) vergeleken met de SOTA-basislijn, wat stabielere temporele dynamiek en scherpere ruimtelijke details aantont. Bovendien geeft de menselijke studie in Tabel 2 aan dat 67,40% van de deelnemers onze video's verkiezen boven concurrerende methoden, wat het perceptuele realisme en de algehele getrouwheid van onze generaties benadrukt.

Kwalitatieve vergelijkingen in Figuur 4 laten verder zien dat Captain Safari visueel aantrekkelijke, realistische en zeer authentieke scène-dynamiek produceert. Deze bevindingen komen ook overeen met de voorbeelden getoond in Figuur 1, waar onze methode levendige, coherente en natuurlijk ogende dronevideo's levert die sterk lijken op opnames uit de echte wereld.

5.4. 3D Consistentie

Captain Safari bereikt state-of-the-art 3D consistentie. Zoals getoond in Tabel 1, verlaagt onze methode MEt3R met 0.0013 (0.3690 vs. 0.3703) en verhoogt het reconstructiepercentage met 0.045 (0.968 vs. 0.923) vergeleken met de sterkste basislijn. Consistent hiermee toont de menselijke studie in Tabel 2 aan dat 65,60% van de deelnemers de voorkeur geeft aan Captain Safari voor 3D consistentie, wat alle concurrerende benaderingen aanzienlijk overtreft.

Kwalitatieve visualisaties bevestigen verder deze kwantitatieve verbeteringen. In Figuur 1 blijven structuren zoals de Griekse zuilen geometrisch stabiel bij grote veranderingen in het gezichtspunt. In Figuur 4 produceert ons model (*links*) een schoolbus die soepel uit het frame beweegt, en (*rechts*) behoudt scherpe, wereldwijd consistentie veldmarkeringen op het voetbalveld, terwijl basislijnen vervormingen en verdwijningen vertonen. Figuur 5 en Figuur 1 laten verder zien dat onze reconstructies scherpere gevels en goed gevormde ramen opleveren.



Figuur 5. **Scenereconstructie en cameratraject.** Met pose-uitgelijnd geheugen reconstrueert *Captain Safari* een goed gestructureerde gebouwgevel (de variant zonder geheugen vervaagt/vervormt het), wat het voordeel van geheugen aantoon. Het behoudt ook fijne details—geparkeerde auto's en de boom op hun daken—die Wan2.2-5B niet kan behouden. Ondertussen volgt Real-CamI2V slechts een kort pad, terwijl *Captain Safari* het volledige traject met stabiele 3D-structuur bestrijkt, wat de uitdaging benadrukt om 3D consistentie te behouden bij snelle bewegingen.

zonder dat de geometrie instort. Samen bevestigen deze resultaten dat het impliciete wereldgeheugen en pose-geconditioneerde retrieval van *CaptainSafari* effectief de onderliggende 3D-wereld stabiliseren bij agressieve camerabewegingen.

5.5. Trajectvolgning

CaptainSafari levert de meest nauwkeurige trajectvolgning van alle concurrerende modellen. Zoals te zien in Tabel 1, behaalt onze methode de hoogste AUC@30 (0.200) en AUC@15 (0.068), samen met de beste cosinusgelijkenis (0.563), waarmee het de sterkste basislijn met duidelijke marges overtreft. De menselijke studie in Tabel 2 versterkt deze observatie verder, met **69,00%** van de deelnemers die ons model als het meest trouw aan het doelcamerapad identificeren.

Figuur 5 biedt een duidelijke visualisatie van deze verbeteringen. *Captain Safari*'s voorspelde traject komt nauw overeen met het werkelijke pad, terwijl de geablateerde variant afwijkt en over het dak vliegt, en RealCam-I2V er niet in slaagt de bedoelde voorwaartse beweging te volgen, slechts licht vooruitgaand in plaats van zich aan het voorgeschreven traject te houden. Bovendien toont onze methode stabiele en coherente generatie onder uitdagende veranderingen in het gezichtspunt met complexe camerabewegingen in Figuur 1. Deze resultaten benadrukken de effectiviteit van ons geheugen-verrijkte, pose-geconditioneerde ontwerp voor nauwkeurige trajecttrouw.

5.6. Ablatie Studie

Onze resultaten benadrukken het belang van het voorgestelde pose-geconditioneerde wereldgeheugen. Zoals getoond in Tabel 1, het toevoegen van geheugen levert aanzienlijke verbeteringen op in zowel 3D-con-

sistentie als trajectvolgning. Deze verbeteringen bevestigen dat het ophalen van pose-uitgelijnde wereldkenmerken bij het doelkader het model een expliciet begrip geeft van hoe de scènes *zou* moeten zijn, wat stabiele geometrie en nauwkeurige bewegingsuitlijning mogelijk maakt.

Kwalitatieve vergelijkingen in Figuur 4 en Figuur 5 illustreren deze effecten verder. Met geheugen behouden de gegenereerde scènes de globale structuur en handhaven ze consistente geometrie over verschillende gezichtspunten. Daarentegen vertoont de geablateerde variant vaak afwijkingen en geometrische inconsistenties. Samen bevestigen deze resultaten de effectiviteit van ons geheugen-verrijkte ontwerp in het stabiliseren van de onderliggende 3D-wereld en het begeleiden van nauwkeurige camerabewegingen.

6. Conclusie

We hebben *Captain Safari* geïntroduceerd, een pose-geconditioneerde wereldmotor gebouwd op een wereldgeheugen dat langeafstands, 3D-consistente videoproduktie mogelijk maakt onder complexe FPV-trajecten. Samen met *OpenSafari*, ons samengestelde dataset van dronevideo's in het wild met geverifieerde cameraposities, vormt dit een rigoureuze benchmark voor controleerbare videoproduktie. *CaptainSafari* verbetert de 3D consistentie en trajectnauwkeurigheid aanzienlijk ten opzichte van eerdere methoden, terwijl het een sterke visuele kwaliteit behoudt. Hoewel het systeem aanzienlijke inferentie-overhead met zich meebrengt, zal toekomstig werk zich richten op real-time wereldmotoren met lichtgewicht geheugen en snellere generatieve basissen. We hopen dat *Captain Safari* en *OpenSafari* verder onderzoek aanmoedigen naar persistente wereldmodellen en lange-termijn controleerbare videoproduktie.

Referenties

- [1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele en Jan Eric Lenssen. Met3r: Het meten van multi-view consistentie in gegenererde beelden. In Proceedings van de Computer Vision and Pattern Recognition Conference, pagina's 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: Traject-geconditioneerde tekst-naar-3d generatie. In Europese Conferentie over Computer Vision, pagina's 53–72. Springer, 2024. 3[3] Shuai Bai, Kegin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technisch rapport. arXiv preprint arXiv:2502.13923, 2025. 6[4] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoufet, Brandon Houghton, Raul Sampedro en Jeff Clune. Video pretraining (vpt): Leren handelen door het bekijken van niet-gelabelde online video's. Voortgang in Neurale Informatie Verwerkende Systemen, 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. Het bakken van Gaussiaans splatten in een diffusie denoiser voor snelle en schaalbare eenstaps beeld-naar-3d generatie en reconstructie. 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: De visie ontmoet drone objectdetectie uitdagingen resultaten. In Proceedings van de IEEE/CVF Internationale Conferentie over Computer Vision, pagina's 2847–2854, 2021. 2[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng en Yinda Zhang. Matterport3d: Leren van rgb-d data in binnenumgevingen. arXiv preprint arXiv:1709.06158, 2017. 1[8] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang en Chuang Gan. AdaWorld: Leren van aanpasbare werelddelen met latente acties. 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: Het controleren van videoproductie met bewegings-trajecten. In Proceedings van de Computer Vision and Pattern Recognition Conference, pagina's 1–12, 2025. 3[10] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li en Ceyuan Yang. CameraCtrl: Het mogelijk maken van camerabesturing voor tekst-naar-video generatie. arXiv preprint arXiv:2404.02101, 2024. 1, 3[11] Chen Hou en Zhibo Chen. Training-vrije camerabesturing voor videoproductie. 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. Lora: Low-rank aanpassing van grote taalmodellen. ICLR, 1(2):3, 2022. 6[13] Ronghang Hu, Nikhila Ravi, Alexander C Berg en Deepak Pathak. Worldsheets: De wereld in een 3D-vel wikkelen voor weergavesynthese vanuit een enkele afbeelding. In Proceedings of the IEEE/CVF Internationale Conferentie over Computer Vision, pagina's 12528–12537, 2021. 1, 2[14] Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He en Li Jiang. Geheugenforcing: Spatio-temporeel geheugen voor consistente scènegenereatie in 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: Langeafstands- en wereldconsistent videodiffusie voor verkenbare 3D-scènegenereatie. arXiv preprint arXiv:2506.04225, 2025. 3[16] Longbin Ji, Lei Zhong, Pengfei Wei en Changjian Li. Pose-traj: Pose-bewuste trajectcontrole in videodiffusie. In Proceedings of the Computer Vision and Pattern Recognition Conference, pagina's 22776–22785, 2025. 1, 2, 3[17] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas en Gordon Wetzstein. Collaboratieve videodiffusie: Consistente multi-videogeneratie met camerabesturing. Voortgang in Neurale Informatie Verwerkende Systemen, 37:16240–16271, 2024. 3[18] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas en Gordon Wetzstein. Collaboratieve videodiffusie: Consistente multi-videogeneratie met camerabesturing. Voortgang in Neurale Informatie Verwerkende Systemen, 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, et al. Realcam-i2v: Real-world afbeelding-naar-video generatie met interactieve complexe camerabesturing. In Proceedings of the IEEE/CVF Internationale Conferentie over Computer Vision, pagina's 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 en Jian Ren. Wonderland: Navigeren door 3D-scènes vanuit een enkele afbeelding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pagina's 798–810, 2025. 1, 3[21] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov en Carl Vondrick. Zero-1-to-3: Zero-shot één afbeelding naar 3D-object. In Proceedings of the IEEE/CVF internationale conferentie over computer vision, pagina's 9298–9309, 2023. 2[22] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang en Li Zhang. Wovogen: Wereldvolume-bewuste diffusie voor controleerbare multi-camera rijscènegenereatie. In Europese Conferentie over Computer Vision, pagina's 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: Genereren van een verkenbare wereld. arXiv preprint arXiv:2412.09624, 2024. 1, 2

- [24] Andrew Melnik, Michal Ljubljjanac, Cong Lu, Qi Yan, Weiming Ren en Helge Ritter. Videodiffusiemodellen: Een overzicht. arXiv preprint arXiv:2405.03150, 2024. 2[25] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan en Jian Zhang. Revideo: Maak een video opnieuw met bewegings- en inhoudscontrole. Voortgang in Neurale Informatie Verwerkende Systemen, 37:18481–18505, 2024. 3[26] Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Weijie Wang, Haoyun Li, Guosheng Zhao, Jie Li, Wenkang Qin, Guan Huang en Wenjun Mei. Wonderturbo: Genereren van een interactieve 3D-wereld in 0,72 seconden. arXiv preprint arXiv:2504.02261, 2025. 1, 3[27] Ava Pun, Gary Sun, Jingkang Wang, Yun Chen, Ze Yang, Sivabalan Manivasagam, Wei-Chiu Ma en Raquel Urtasun. Neurale lichtsimulatie voor stedelijke scènes. Voortgang in Neurale Informatie Verwerkende Systemen, 36 19326, 2023 19291 2 : – .
- [28] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3D dataset (hm3d): 1000 grootschalige 3D-omgevingen voor belichaamde AI. arXiv preprint arXiv:2109.08238, 2021. 1[29] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas M'uller, Alexander Keller, Sanja Fidler en Jun Gao. Gen3c: 3D-geïnformeerde wereldconsistente videogeneratie met precieze camerabesturing. In Proceedings of the Computer Vision and Pattern Recognition Conference, pagina's 6121–6132, 2025. 1, 3[30] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart en Marcin Dymczyk. Van grof naar fijn: Robuuste hiërarchische lokalisatie op grote schaal. In CVPR, 2019. 5, 7[31] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz en Andrew Rabinovich. SuperGlue: Leren van kenmerkmatching met grafische neurale netwerken. In CVPR, 2020. 5, 7 [32] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: Een platform voor onderzoek naar belichaamde AI. In Proceedings of the IEEE/CVF internationale conferentie over computer vision, pagina's 9339–9347, 2019. 1[33] Manuel-Andreas Schneider, Lukas Höllein en Matthias Nießner. Worldexplorer: Naar het genereren van volledig navigeerbare 3D-scènes. arXiv preprint arXiv:2506.01799, 2025. 3[34] Xincheng Shuai, Henghui Ding, Zhenyuan Qin, Hao Luo, Xingjun Ma en Dacheng Tao. Vrije-vorm bewegingscontrole: Beheersen van de 6D-posities van camera en objecten in videogeneratie. In Proceedings of the IEEE/CVF Internationale Conferentie over Computer Vision, pagina's 12449–12458, 2025. 2[35] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren en Raquel Urtasun. Scenegen: Leren om realistische verkeersscènes te genereren. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pagina's 892–901, 2021. 2[36] Zachary Teed en Jia Deng. RAFT: Recurrent all-pairs field transforms voor optische stroom. In Europese conferentie over computer vision, pagina's 402–419. Springer, 2020. 5
- [37] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, en Sylvain Gelly. Naar nauwkeurige generatieve modellen van video: Een nieuwe metriek & uitdagingen. arXiv preprint arXiv:1812.01717, 2018. 7[38] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jin-gren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yi-tong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, en Ziyu Liu. Wan: Open en geavanceerde grootschalige generatieve videomodellen. arXiv preprint arXiv:2503.20314, 2025. 6, 7[39] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, en David Novotny. Vggt: Visuele geometrie gegrondeerde transformer. In Proceedings van de Computer Vision and Pattern Recognition Conference, pagina's 5294–5306, 2025. 7[40] Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahao Zhang, Yuxiang Guo, Xijun Liu, Rama Chellappa, Cheng Peng, Alan Yuille, et al. EvoWorld: Evoluerende panoramische wereldgeneratie met expliciete 3D-geheugen. arXiv preprint arXiv:2510.01183, 2025. 1, 2, 3[41] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiumei Wang, Yingya Zhang, Yujun Shen, Deli Zhao, en Jingren Zhou. Videocomposer: Compositorische videosynthese met bewegingscontroleerbaarheid. Voortgang in Neurale Informatie Verwerkende Systemen, 36:7594–7611, 2023. 3[42] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshu Chen, Menghan Xia, Ping Luo, en Ying Shan. Motionctrl: Een verenigde en flexibele bewegingscontroller voor videogeneratie. In ACM SIGGRAPH 2024 Conference Papers, pagina's 1–11, 2024. 3[43] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, en Justin Johnson. Synsin: End-to-end weergavesynthese vanuit een enkele afbeelding. In Proceedings van de IEEE/CVF-conferentie over computervisie en patroonherkenning, pagina's 7467–7477, 2020. 1, 2 – . [44] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, en Jiang Bian. Geometrie Dwang: Het combineren van videodiffusie en 3D-representatie voor consistent wereldmodellering. arXiv preprint arXiv:2507.07982, 2025. 3, 6, 7[45] Jianzong Wu, Liang Hou, Haotian Yang, Xin Tao, Ye Tian, Pengfei Wan, Di Zhang, en Yunhai Tong. Vmob: Mengsel-van-blok aandacht voor videodiffusiemodellen. arXiv preprint arXiv:2506.23858, 2025. 1, 2[46] Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gerneth, Liangliang Cao, Zhangyang Wang, en Hao Tang. Cavia: Camera-controleerbare multi-view videodiffusie met weergave-geïntegreerde aandacht. arXiv preprint arXiv:2410.10774, 2024. 3

- [47] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, en Arash Vahdat. Camco: Camera-gestuurde 3D-consistente afbeelding-naar-video generatie. arXiv preprint arXiv:2406.02509, 2024. 3
- [48] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, en 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: Geünificeerde multimodale video generatie via camera controle. arXiv preprint arXiv:2504.02312, 2025. 2
- [50] Alex Yu, Vickie Ye, Matthew Tancik, en Angjoo Kanazawa. pixelnerf: Neurale stralingsvelden van één of enkele afbeeldingen. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pagina's 4578–4587, 2021. 1, 2
- [51] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, en Jiajun Wu. Wonderworld: Interactieve 3D scène generatie vanuit een enkele afbeelding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pagina's 5916–5926, 2025. 1, 3
- [52] Mark YU, Wenbo Hu, Jinbo Xing, en Ying Shan. Trajectorycrafter: Het herleiden van camera trajecten voor monocular video's via diffusiemodellen. 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, en Yonghong Tian. Viewcrafter: Het temmen van videodiffusiemodellen voor hoogwaardig nieuwe weergave synthese. arXiv preprint arXiv:2409.02048, 2024. 3
- [54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, en Oliver Wang. De onredelijke effectiviteit van diepe kenmerken als een perceptuele maatstaf. In Proceedings of the IEEE conference on computer vision and pattern recognition, pagina's 586–595, 2018. 7
- [55] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, en Weizhi Wang. Tora: Trajectorie-georiënteerde diffusie transformator voor video generatie. In Proceedings of the Computer Vision and Pattern Recognition Conference, pagina's 2063–2073, 2025. 3
- [56] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, en Tao Mei. Motionpro: Een precieze bewegingscontroller voor afbeelding-naar-video generatie. In Proceedings of the Computer Vision and Pattern Recognition Conference, pagina's 27957–27967, 2025. 3
- [57] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, en Xi Li. Cami2v: Camera-gestuurd afbeelding-naar-video diffusiemodel. arXiv preprint arXiv:2410.15957, 2024. 1, 3
- [58] Mengqi Zhou, Yuxi Wang, Jun Hou, Shougao Zhang, Yiwei Li, Chuanchen Luo, Junran Peng, en Zhaoxiang Zhang. Scenex: Procedureel controleerbare grootschalige scène generatie. arXiv preprint arXiv:2403.15698, 2024. 2
- [59] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, en Noah Snavely. Stereo vergroting: Leren van weergave synthese met behulp van multiplane afbeeldingen. arXiv preprint arXiv:1805.09817, 2018. 2, 5
- [60] Yunsong Zhou, Michael Simon, Zhenghao Peng, Sicheng Mo, Hongzi Zhu, Minyi Guo, en Bolei Zhou. Simgen: Simulator-geconditioneerde generatie van rijscènes. *Voortgang in Neurale Informatie Verwerkende Systemen*, 3748874, 202448838:–.
- [61] Zhenghong Zhou, Jie An, en Jiebo Luo. Latent-reframe: Camera controle mogelijk maken voor videodiffusiemodellen zonder training. In *Proceedings of the IEEE/CVF Internationale Conferentie over Computer Vision*, pagina's 12779–12789, 2025. 3
- [62] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, en Jiwen Lu. Streaming 4d visuele geometrie transformator. arXiv preprint arXiv:2507.11539, 2025. 6