

Experiments in Gaussian Splatting for VFX Scene Reconstruction

Chen Wu
Amazon Web Services
Australia
wuc@amazon.com

GG Heitmann
Wētā FX
New Zealand
gg@wetafx.co.nz

Kimball Thurston
Wētā FX
New Zealand
kthurston@wetafx.co.nz

Yin Song
Amazon Web Services
Australia
yinsong@amazon.com

Peter Hillman
Wētā FX
New Zealand
peterh@wetafx.co.nz

Josh Passenger
Amazon Web Services
Australia
jospas@amazon.com

K. Ho
Amazon Web Services
Australia

Ibrahim Gabr
Wētā FX
New Zealand
igabr@wetafx.co.nz

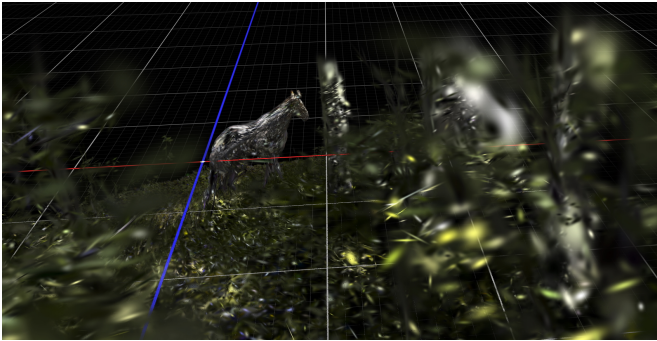
Viral Shah
Amazon Web Services
New Zealand
vvshah@amazon.com



(a) Ground truth (Input frames)



(b) 4DGS-SOM @ 21.56 PSNR



(c) Novel view 1 from SuperSplat



(d) Novel view 2 from SuperSplat

Figure 1: Reconstruction results for the horseInWoods_v02 scene. Top: (a) ground truth (b) our 4DGS-SOM method achieving 21.56 PSNR. Bottom: (c, d) additional novel viewpoints rendered from the 4DGS-SOM reconstruction. (c) is rendered from the opposite side of the shooting camera, showing the far side of the horse that was never observed during training.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DigiPro '26, July 18, 2026, Los Angeles, CA, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN XXXX-XXXX-X/26/08
<https://doi.org/XXXXXXX.XXXXXXX>

ABSTRACT

We describe a series of experiments reconstructing dynamic VFX scenes from multi-camera and moving-camera footage using Gaussian Splatting, built from open-source components. Through this process we report what worked, what broke, and what we learned, spanning an extension of Shape of Motion for moving cameras, an evaluation harness for 4D Gaussian Splatting, a baseline study, and algorithmic contributions targeting core limitations of existing methods. We report results on multi-camera performance-capture data from visual effects production, showing freely navigable 4D scene reconstructions.

CCS CONCEPTS

• **Computing methodologies** → **Image-based rendering**; *Image manipulation*.

KEYWORDS

Gaussian Splatting, 4D Reconstruction, Visual Effects, Novel View Synthesis, Depth Estimation

ACM Reference Format:

Chen Wu, Yin Song, K. Ho, GG Heitmann, Peter Hillman, Ibrahim Gabr, Kimball Thurston, Josh Passenger, and Viral Shah. 2026. Experiments in Gaussian Splatting for VFX Scene Reconstruction. In *Proceedings of the Digital Production Symposium 2026 (DigiPro '26)*. ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Layering CG elements with live action plates is one of the most challenging tasks in Visual Effects. Deep Compositing techniques allow accurate combination of different CG elements, even when the elements need to be interleaved together, but such techniques rely on extra information provided by the renderer, which is not available for live action plates. Increasingly, Machine Learning algorithms are applied to plates and other captured footage to estimate auxiliary information such as depth maps, segmentation masks, rotoscoping splines, and motion vectors. This data is an invaluable aid to visual effects artists working on shots, but it remains fundamentally 2D.

3D Gaussian Splats [Kerbl et al. 2023] (3DGS) are analogous to deep images, and offer a promising approach for compositing. Gaussian Splats store scenes as clouds of independently colored volumetric primitives that are fast to render and compatible with existing deep compositing workflows: CG deep images can be merged with Gaussian Splats to achieve correct layering.

Robust tools exist for generating Gaussian Splats of static objects such as props or sets. These require hundreds of still images or video taken with a very slow moving camera. Extending this to typical production conditions scenes is the missing piece: production footage has moving actors, a limited number of arbitrarily moving cameras, with limited opportunities for additional cameras and sensors. Published methods for generating splats do not yet handle these conditions reliably.

Traditional approaches such as LIDAR scanning capture accurate static set geometry but require dedicated hardware, provide no appearance model, and operate too slowly to capture motion. 3DGS reconstructs both geometry and appearance from passive camera

imagery that production is already capturing, including dynamic scenes with moving subjects.

For a Gaussian Splat representation of a plate to be practical, it must be as close to the plate as possible, ideally including effects such as sensor noise, and depth of field and motion blur. In practice, the original plate can be used in place of the splats for much of the image, as long as the result is close enough.

In this paper, we describe our experiments building a family of Gaussian Splatting pipelines for reconstructing dynamic scenes from multi-camera and moving-camera footage, built from open-source components. Our source footage came from scenes captured by up to ten cameras, shooting production resolution with a variety of camera types and per-camera exposure variation. A single camera synthetic tracking shot was rendered to provide a source with ground truth comparison. We begin in Section 2 with a per-frame 3DGS baseline study that establishes the empirical foundation for our work. In Section 3 we discuss our implementation of a recent 4DGS method [Cho et al. 2025; Wu et al. 2024] on our production footage. In Section 4 we present 4DGS-SOM, our contribution targeting the core limitation of existing methods such as [Wang et al. 2025].

2 PER-FRAME 3DGS BASELINE

The datasets for our experiments comprise four multi-camera sequences of actors walking around in a studio (“milling” sequences) and one single-camera synthetic scene of a horse in a forest (“horse-in-woods” sequence). The multi-camera datasets used eight Sony hero cameras for training with two RED cameras held out for novel-view evaluation. We began with a per-frame 3DGS baseline across these five datasets, establishing the empirical foundation for the pipelines that follow. Each output frame is a standalone 3DGS model, making the system easy to inspect and debug. The pipeline decomposes reconstruction into three stages: data preparation, monocular metric depth caching, and per-frame 3DGS training with configurable depth supervision. We evaluated 13 configurations along three axes (chunk size, depth weight, and context-view weight) at reduced fidelity (5,000 iterations, half resolution), then validated the top candidates at full resolution (30,000 iterations). These experiments revealed some early lessons:

- Without per-camera exposure correction, gain and white-balance differences between sensors introduce up to 4 dB of artificial degradation in held-out camera evaluation.
- Initialisation quality strongly affects reconstruction quality.
- Extended training schedules yielded diminishing returns, contributing less than 0.5 dB.

These lessons informed the exploration of the two 4DGS methods that follow: an evaluation harness for the 4D-Scaffold-GS method and our own approach 4DGS-SOM to deal with dynamic scenes with a moving camera.

3 4D-SCAFFOLD-GS

4D-Scaffold-GS [Cho et al. 2025] extends 3DGS to dynamic scenes by anchoring Gaussians to a deformable scaffold and predicting time-varying attributes through Multi-Layer Perceptrons (MLPs). We implemented an evaluation harness for this multi-view 4DGS method and applied it to our own synthetic and live-action footage.

Using this harness, we explored the impact of several training pipeline design decisions on 4D reconstruction accuracy and efficiency. Throughout the development process, we generated a series of benchmarks not only for performance tracking but also for debugging purposes, encountering and solving for a series of technical issues common when implementing bleeding-edge methods from the literature to industrial use cases. These issues stemmed from differences in 3D coordinate conventions, the importance of Gaussian point initialisation, and compute scalability. This required the development of diagnostic tools for splat and frustum visualisation, which were incorporated into the 4D-Scaffold-GS training loop. Our efforts into 4D-Scaffold-GS have yielded a reliable platform for the future development and evaluation of methods for novel view synthesis in motion-picture VFX post-production.

Table 1 summarizes the 4D-Scaffold-GS reconstruction quality across four multi-camera production datasets, distinct from the single-camera horseInWoods_v02 scene shown in Figure 5. These datasets feature indoor captures of four characters milling around (walking, waving hands) against blue screen backgrounds with empty floors; datasets labeled “millsetclean” include additional scene complexity such as chairs, reflective glasses, containers, and mesh objects. Training PSNR remains consistently high across all sequences, while novel view synthesis quality on hold-out cameras varies significantly by dataset and viewpoint. The large gap between training and hold-out PSNR is partially attributable to rig sparsity and partially to uncorrected gain and white-balance differences between the Sony training cameras and RED evaluation cameras, as identified in our baseline study (Section 2).

Table 1: 4D-Scaffold-GS mean PSNR on training and hold-out cameras across four “milling” sequence datasets (120K iterations).

Dataset	Train PSNR	Hold-out Cam 1	Hold-out Cam 2
millbclean_seq1	34.48	8.90	10.84
millbclean_seq2	34.06	8.54	11.07
millsetbclean_seq1	32.80	10.65	—
millsetbclean_seq2	32.24	12.02	—

4DGS methods typically struggle with non-linear fast movement because each Gaussian follows a linear trajectory and requires a narrow temporal visibility window to avoid ghosting [Gyeon Kim et al. 2026], as illustrated in Figure 2. We observed that 4D-Scaffold-GS can approximate mild view-dependent effects (soft specular highlights) through its view-conditioned color MLP, and handles fine volumetric structures (netting, foliage) given sufficient anchor density, as shown in Figure 3.

To observe the 4D-Scaffold-GS performance on novel view synthesis, we held-out camera views from the “millbclean” and “millsetbclean” datasets. Figure 4 shows a representative example of novel view synthesis quality on the held-out camera for the millsetbclean_seq2 dataset. In this dataset, the camera crew member (far left in Figure a) and equipment are not seen in any of the training camera views and, as expected, are unable to be reconstructed in the novel view. Conversely, it is encouraging that the actors and props



Figure 2: Fast-movement, prop-free reconstruction on millbclean_seq2. The actor’s right hand disappears mid-star-jump, a characteristic failure of linear per-Gaussian motion under non-linear, fast trajectories with limited temporal supervision.

blocked by equipment in the ground-truth views are reconstructed and remain recognizable, albeit at low fidelity.

The horseInWoods_v02 scene poses the greatest challenge as it features a single moving camera filming a moving subject. Scenes with more dynamic regions require the growth of more anchors during training, subsequently increasing the resulting model size and the required training time [Cho et al. 2025], as shown in Figure 5. While known camera view reconstructions even in this use-case still produces high quality results, the single-camera case also exposes *depth ambiguity*: 4D-Scaffold-GS cannot uniquely determine the true depth of scene elements from one viewpoint. This limitation motivates our own approach, presented in the next section.

4 4DGS-SOM

The depth ambiguity exposed in Section 3 stems from a lack of effective multi-view constraint on the static background. Like our work, Shape of Motion [Wang et al. 2025] factors the scene into a dynamic foreground (with low-rank $\mathbb{SE}(3)$ motion bases) and a static background. However, it trains both halves with a single monocular depth source and applies its 2D track-consistency loss only to the foreground. Under significant camera motion this produces visible artifacts: backgrounds that slide or drift toward the camera, foreground edges that disintegrate over time, and halos around subject silhouettes.

To tackle this issue, we developed 4DGS-SOM as shown in Figure 6, in which the foreground retains monocular per-frame depth supervision (Depth Anything V1 [Yang et al. 2024], metric-aligned via Metric3D [Hu et al. 2024]). However, the background is instead supervised by *multi-view-consistent* depth from Depth Anything 3 [Lin et al. 2026], which exploits static background parallax across frames. This is something a per-frame monocular depth estimation method like DA-1 cannot exploit. In particular, we recover background 3D positions by classical multi-view triangulation. BootsTAPIR [Doersch et al. 2024] 2D tracks across frames are treated as observations of static 3D points seen from many cameras. A geometric-conditioning filter rejects ill-posed cases (low parallax, near-parallel viewing rays) before the linear solve. This produces seed positions wherever parallax is sufficient.

We then make two changes to Shape of Motion’s loss functions, each targeting a failure mode that pixel-wise photometric loss cannot resolve. **Full-distribution mask loss.** Shape of Motion applies a per-pixel L1 between rendered alpha and binary mask, trimming



Figure 3: Fast-movement reconstruction with props on millsetbclean_seq2. 4D-Scaffold-GS can approximate mild view-dependent effects (soft specular highlights) through its view-conditioned color MLP, and handles fine volumetric structures (netting, foliage) given sufficient anchor density.

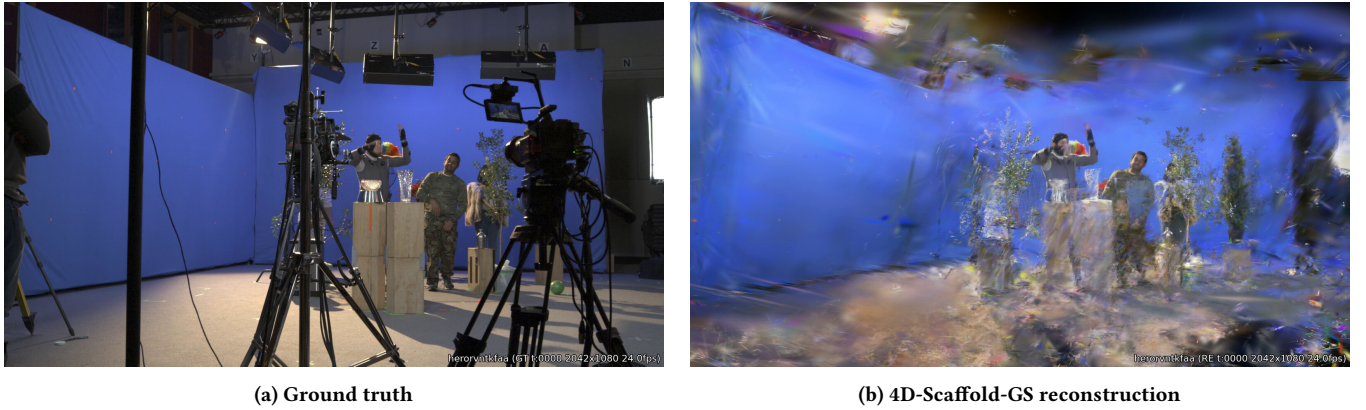


Figure 4: Novel view synthesis on held-out camera from millsetbclean_seq2, demonstrating the quality of view synthesis on cameras and crewmember not seen during training.



Figure 5: The reconstructions of the horseInWoods_v02 scene using 4D-Scaffold-GS [Cho et al. 2025] at (b) 60K and (c) 90K iterations compared to (a) the ground truth.

the top 2% of errors for robustness. However, rendered alpha decays softly at Gaussian edges while the ground-truth mask is binary, so boundary pixels always carry the largest error and always fall in the trimmed tail. The result is zero gradient at silhouette edges, producing visible halos. We retain the full error distribution, restoring

gradient at boundaries and eliminating the artifacts. **Background 2D track-consistency loss.** Shape of Motion applies its track residual loss only to foreground Gaussians. We extend the same loss to static background Gaussians, with tracks recomputed on

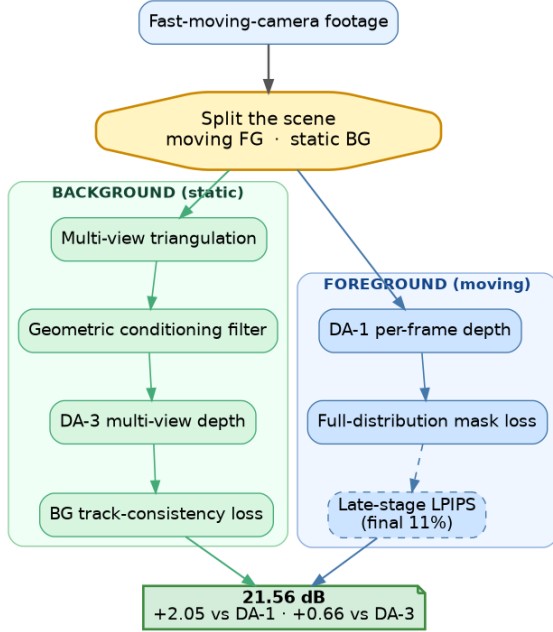


Figure 6: An overview of our 4DGS-SOM method.

the background mask region. Without this loss, background Gaussians elongate along the camera ray over long training schedules, producing streaked artifacts in novel views.

The result is promising as shown in Figure 1(b, c, d). On a 70-frame moving-camera sequence, static environments gain realistic layered 3D structure (objects at correct distances, not flattened), and moving subjects keep sharp, stable silhouettes under apparent camera motion.

4.1 Shape of Motion vs. 4DGS-SOM

We compare our reconstruction quality against the upstream Shape of Motion baseline. Both methods are trained and evaluated on all 70 frames of the input sequence, and their results are summarized in Table 2. Our method has improved the mean PSNR by +2.80 dB over the baseline, with consistent gains across all four statistics.

Table 2: PSNR comparison between our 4DGS-SOM and Shape of Motion baseline on the horseInWoods_v02 scene (70 frames).

Method	Mean PSNR
Shape of Motion (baseline)	18.76
4DGS-SOM (Ours)	21.56
Improvement	+2.80

Moreover, in the Shape of Motion baseline, the background is modelled by static Gaussians supervised only by per-frame photometric loss. From a single viewpoint, this loss provides no gradient along the viewing ray, leaving depth along the camera ray unobservable. With limited camera motion, this ambiguity is hidden, but

under significant camera motion it manifests as streaking artefacts in novel views as shown in Figure 7.



Figure 7: Streaking background artefacts in the Shape of Motion baseline. Significant camera motion creates many distinct viewpoints of the static background, but the per-frame photometric loss fits each viewpoint independently with no multi-view consistency constraint. Depth along the camera ray is therefore under-constrained, and over training the background Gaussians drift along their view rays and stretch into thin ribbons.

4.2 4D-Scaffold-GS vs. 4DGS-SOM

A characteristic limitation of monocular view 4D-Scaffold-GS reconstruction is depth ambiguity. When reconstructing from a single viewpoint, the method cannot uniquely determine the true depth of scene elements, leading to reconstructions where objects are represented by multiple layers of Gaussians at varying depths along the viewing ray, as illustrated in Figure 8a. This parallax error is inherent to monocular reconstruction and manifests as ghosting and layering artifacts in novel views. This remains a fundamental challenge for single-camera 4DGS applications.

4DGS-SOM addresses this challenge by augmenting the initial point cloud with information from separate depth estimators. As Figure 8 shows, augmenting the static background with multi-view depth supervision resolves the layered ambiguity that conventional single-camera 4D-Scaffold-GS leaves behind. This gain is not captured by PSNR alone: Table 3 places all three of our methods on the common horseInWoods_v02 scene — per-frame 3DGS at 17.75, conventional 4D-Scaffold-GS at 21.54, and 4DGS-SOM at 21.56 — where conventional 4D-Scaffold-GS and 4DGS-SOM are nearly indistinguishable despite the clear qualitative difference.

Table 3: Mean PSNR of 3DGS, 4D-Scaffold-GS, and 4DGS-SOM on the horseInWoods_v02 scene (single-camera).

Method	Mean PSNR
3DGS	17.75
4D-Scaffold-GS	21.54
4DGS-SOM	21.56

5 SPLAT APPLICATIONS

Figure 9 shows a use of splats to create a novel viewpoint close to the original. A 5 degree shift from the original position introduces

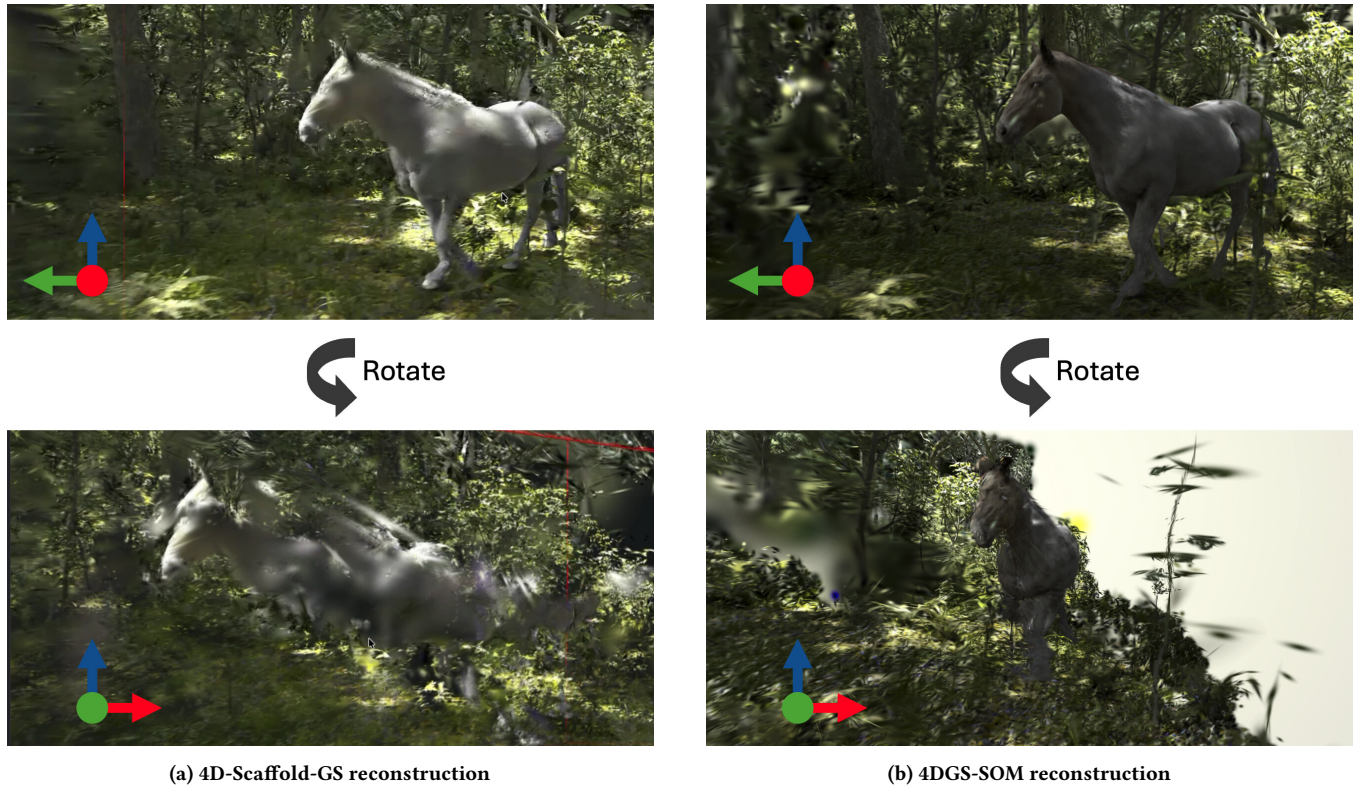


Figure 8: Examples of how the 4DGS-SOM method addresses the depth ambiguity seen with conventional monocular 4DGS reconstruction. (a) shows the 4D-Scaffold-GS reconstruction exhibiting multiple layers of Gaussians at varying depths along the viewing ray, while (b) demonstrates how 4DGS-SOM captures the horse volumetrically as a coherent 3D structure, mitigating the depth ambiguity through its static background modeling approach.

only minimal artefacts. Increasing the shift to 10 degrees creates unacceptably high distortions. Small offsets like this would be typical when modifying a plate to stabilize the camera, or to align two passes.

Figure 10 shows a use of our splat system for compositing tasks: isolating the horse from the background, for example in order to insert an object into the scene. Where only an alpha channel is available, edges are susceptible to contamination or “bleed through” as the color contribution of the background cannot be removed completely from the foreground image. This is most apparent in the area between the horse’s front legs. Although the splats have not fully recovered all the edge detail and fine hairs, they have successfully separated the color of the horse from the background, so there is no bleed through. Figure 10(d) shows the splats scene merged with the Cactus library image from Steam Studio. The edge detail appears correct. This image was created simply by transforming and combining the splats: it was not necessary to split the splats into pieces to achieve the correct result.

6 LIMITATIONS AND FUTURE WORK

Starting from a simple per-frame 3DGS baseline and progressing to a 4DGS implementation, we have developed techniques for reconstructing production VFX footage as freely navigable volumetric

scenes. Although the quality is already useful as a visual reference during production, our reconstructions are not high enough quality for final pixels. For example, novel views still show visible artifacts, particularly around fine detail and object boundaries. Motion blur also proved challenging to handle. For a more robust evaluation of our technique, a wider variety of sources would be required, including live action shots with moving cameras, and scenes with more complex types of background motion.

Our 4DGS-SOM method assumes a clean foreground / static-background split, which could struggle with autonomous background motion such as crashing waves or wind-driven foliage. Both Shape of Motion and our extension model the background as static. Addressing this requires giving background Gaussians their own low-rank motion bases, separate from those in the foreground. These would be trained by the multi-view 2D track-consistency loss we already apply for background correspondence. The trainer could then fit autonomous background motion. We leave validation of this proposal to our future work.

Looking ahead, we see several directions that could close the remaining quality gap. Hybrid approaches that combine Gaussian Splats with traditional meshes, memory-efficient representations for longer sequences, and material-aware Gaussians for relighting are all promising directions. Of these, we believe foundation depth

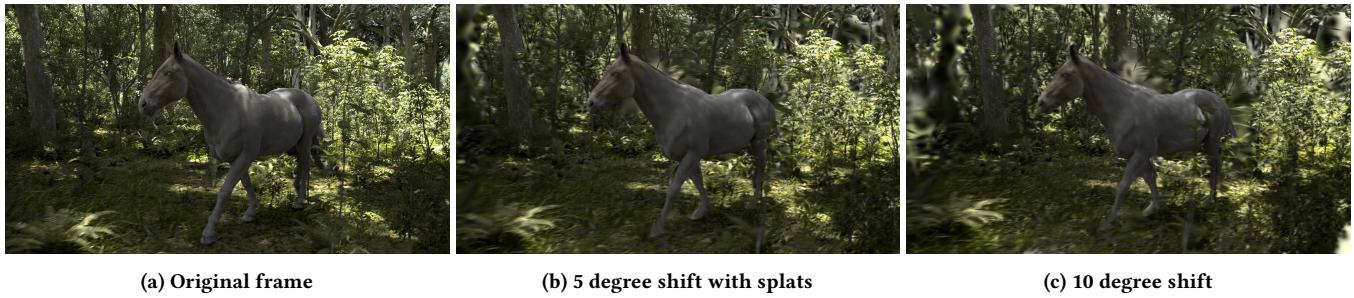


Figure 9: Using splats to create a novel viewpoint. Left to right: original frame, 5 degree shift with splats, 10 degree shift.

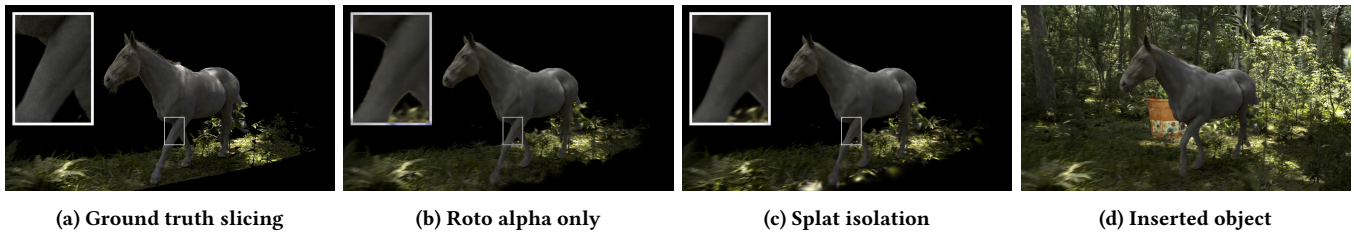


Figure 10: Using splats to isolate parts of a scene. (a) ground truth slicing obtained from the rendered deep image; (b) edge contamination around horses legs when using roto for alpha alone; (c) better isolation with splats; (d) example application inserting a new object into the scene.

models will have the greatest impact. Our 4DGS-SOM results show that reconstruction quality is bounded by the accuracy of depth estimation for both foreground and background branches. As these models produce more accurate metric-scale estimates from fewer input views, our reconstruction quality improves directly.

By exploiting foundation depth models (e.g. [Gabeur et al. 2026]) within the Gaussian Splatting framework, our pipeline benefits directly from the scaling laws driving the broader AI industry. As foundation depth models continue to improve rapidly, so does our reconstruction quality since our pipelines are already structured to integrate these improvements as they arrive. More broadly, the methods and tooling described in this talk are not tied to a single architecture or dataset. They address the practical gap between published Gaussian Splatting research and production capture requirements, and we hope our work can help accelerate the adoption of Gaussian Splatting methods in production VFX.

ACKNOWLEDGMENTS

We thank the teams at Amazon Web Services and Wētā FX for their support of this research collaboration.

REFERENCES

- Woong Oh Cho, In Cho, Seoha Kim, Jeongmin Bae, Youngjung Uh, and Seon Joo Kim. 2025. 4D Scaffold Gaussian Splatting with Dynamic-Aware Anchor Growing for Efficient and High-Fidelity Dynamic Scene Reconstruction. In *Arxiv*.
- Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, and Andrew Zisserman. 2024. BootsTAP: Bootstrapped Training for Tracking-Any-Point. In *Asian Conference on Computer Vision (ACCV)*.
- Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. 2026. Image Generators are Generalist Vision Learners. *arXiv preprint arXiv:2604.20329* (2026).
- Seung Gyeom Kim, Areum Kim, Yongjae Yoo, and Sukmin Yun. 2026. Implicit 4D Gaussian Splatting for Fast Motion with Large Inter-Frame Displacements. In *The Fourteenth International Conference on Learning Representations*.
- Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. 2024. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 10579–10596.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 42, 4 (2023).
- Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. 2026. Depth anything 3: Recovering the visual space from any views. *ICLR 2026 Oral* (2026).
- Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. 2025. Shape of motion: 4d reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9660–9672.
- Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20310–20320.
- Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10371–10381.