

4DGT: Learning a 4D Gaussian Transformer Using Real-World Monocular Videos

Zhen Xu^{1,2,*} Zhengqin Li¹ Zhao Dong¹ Xiaowei Zhou² Richard Newcombe¹ Zhaoyang Lv¹

¹Reality Labs Research, Meta ²Zhejiang University

Project page: <https://4dgt.github.io>

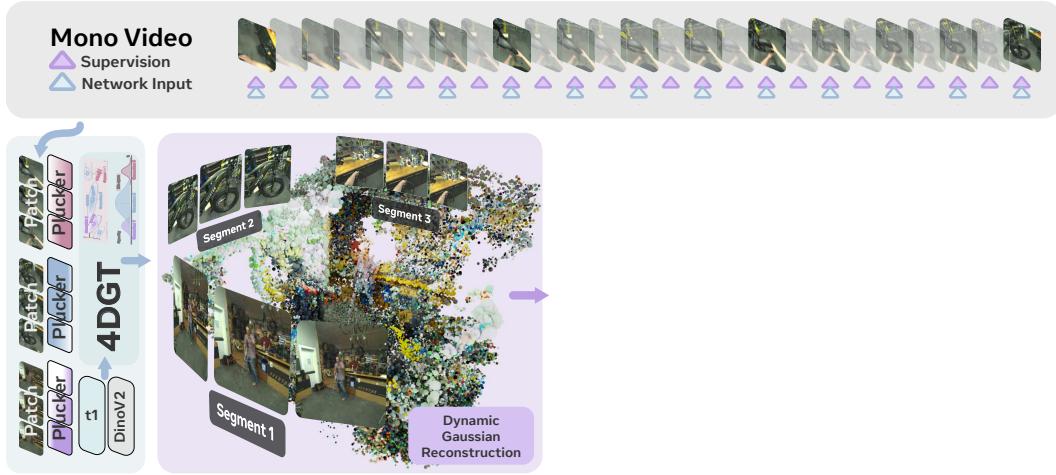


Figure 1: We propose a scalable 4D dynamic reconstruction model trained only on real-world monocular RGB videos. The feed-forward 4DGs (section 3.1) representation enables us to render the geometry and appearance of the dynamic scene from novel views in real-time. Even without explicit supervision, the model can learn to distinguish dynamic contents from the background and produce realistic optical flows. The figure shows an enlarged set of Gaussians for the purpose of visualization. **The embedded rendered videos only play in Adobe Reader or KDE Okular.**

Abstract

We propose 4DGT, a 4D Gaussian-based Transformer model for dynamic scene reconstruction, trained entirely on real-world monocular posed videos. Using 4D Gaussian as an inductive bias, 4DGT unifies static and dynamic components, enabling the modeling of complex, time-varying environments with varying object lifespans. We proposed a novel density control strategy in training which enables our 4DGT to handle longer space-time input and remain efficient rendering at run-time. Our model processes 64 consecutive posed frames in a rolling-window fashion, predicting consistent 4D Gaussians in the scene. Unlike optimization-based methods, 4DGT performs purely feed-forward inference, reducing reconstruction time from hours to seconds and scaling effectively to long video sequences. Trained only on large-scale monocular posed video datasets, 4DGT can outperform prior Gaussian-based networks significantly in real-world videos and achieve on-par accuracy with optimization-based methods on cross-domain videos.

* Work done during internship at Meta.

1 Introduction

Humans record videos to digitize interactions with their surroundings. The ability to recover persistent geometry and 4D motion from videos has a profound impact on AR/VR, robotics, and content creation. Modeling 4D dynamic interactions from general-purpose videos remains a long-standing challenge. Prior work relying on multi-view synchronized capture or depth sensing is constrained to specific application domains. Recent progress in monocular dynamic video reconstruction via per-video optimization shows promise, but lacks scalability due to its time-consuming inference.

In this paper, we propose 4D Gaussian Transformer (4DGT), a novel transformer-based model that reconstructs dynamic scenes from posed monocular videos in a feedforward manner. We assume camera calibration and 6-degree-of-freedom (6DoF) poses are available from on-device SLAM [9] or offline pipelines [23, 31]. Inspired by recent feedforward reconstruction methods for static 3D scenes [67, 70], 4DGT learns reconstruction from data and adopts 4D Gaussian Splatting (4DGS) [53] as a unified representation for both static and dynamic content, differing only in lifespan. This design enables fast 4D reconstruction from short videos in seconds. For longer videos with global pose consistency, 4DGT predicts consistent world-aligned 4DGS using 64-frame rolling windows.

Training a 4D representation is challenging in defining appropriate supervision. While dynamic monocular videos are abundant, they lack space-time constraints. Multi-view video datasets [4, 22] are limited in both quantity and diversity, making them insufficient for training models that generalize in the wild. Prior methods [42] trained on synthetic object-level data suffer from a generalization gap when applied to complex real-world dynamics.

To address this, we train 4DGT exclusively on posed monocular videos from public datasets. We use two key strategies to mitigate space-time ambiguity. First, we leverage depth and normal predictions from expert models [40, 62, 64] as auxiliary supervision for guiding geometry learning. Second, we regularize predicted Gaussian properties to favor longer lifespans and reduce overfitting to specific views. These enable 4DGT to effectively disentangle space-time structure, yielding high-quality geometry, novel view synthesis, and emergent motion properties such as segmentation and flow. Our reconstructions also show better metric consistency than the expert models used for supervision.

Scaling a transformer to predict dense, pixel-aligned 4DGS presents two main challenges. First, dense pixel-aligned 4DGS predictions are computationally expensive for training and rendering. Inspired by density control in 3DGS [19], we introduce a pruning strategy that removes the redundant pixel-aligned 4DGS and further increases tokens in a second training with denser space-time samples. This effectively reduces 80% of Gaussians and enables a $16\times$ higher sampling rate with the same compute. Second, as space-time samples increase, the number of tokens grows, and vanilla self-attention scales quadratically. To address this, we propose a level-of-detail structure via multi-level spatiotemporal attention, achieving an additional $2\times$ reduction in computational cost.

4DGT is the first transformer-based method for predicting 4DGS in a feedforward manner using only real-world posed monocular videos in training. Across varying scales, it generalizes well to unseen videos from similar device domains. Extensive evaluations across datasets and domains show that 4DGT achieves comparable reconstruction quality to optimization-based methods while being three orders of magnitude faster, making it practical for long video reconstruction. Compared to prior methods that also train dynamic Gaussian prediction [42, 25], 4DGT generalizes better to complex real-world dynamics and also exhibits emergent motion properties such as motion segmentation.

In summary, we make the following technical contributions:

- We introduce 4DGT, a novel 4DGS transformer trained on posed monocular videos at scale which produces consistent 4D video reconstructions in seconds at inference.
- We propose a training strategy to densify and prune space-time pixel-aligned Gaussians, reducing 80% of predictions, achieving $16\times$ higher sampling rate during training and a $5\times$ speed-up in rendering.
- We design a multi-level attention module to efficiently fusing space-time tokens, further reducing training time by half.
- Our experiments demonstrate strong scalability of 4DGT across real-world domains using mixed training datasets and can outperform the previous Gaussian network significantly. The performance of 4DGT is on par with optimization-based methods in accuracy in cross-domain videos recorded by similar devices used in training, while being 3 orders of magnitude faster.

2 Related Work

Nonrigid reconstruction. Recovering dynamic content from video has long been a holy grail challenge in 3D vision. Early approaches demonstrated promising non-rigid shape reconstruction from RGB-D videos [32, 46, 3], but relied heavily on depth input and struggled with complex dynamic scenes. Since the seminal work on Neural Radiance Fields (NeRF) [30], several methods have extended NeRF to 4D using multi-view videos [22, 10] or posed monocular videos [36, 37]. However, 4D NeRFs are slow to train and render, limiting their scalability for complex dynamic scenes. Recently, generative priors have shown promise in aiding 4D reconstruction [56], offering strong regularization for shape and motion across space and time. Still, optimizing 4D representations remains time-consuming, and reconstruction quality depends heavily on the generalizability of priors across different scene domains.

Dynamic Gaussian representations. Since the introduction of 3D Gaussian Splatting [19], several methods have extended it to 4DGS variants [63, 60, 53, 8], showing promising dynamic scene reconstruction from multi-view videos with faster training and real-time rendering support. However, optimizing dynamic Gaussians for monocular videos remains challenging. Recently, a few works have shown that complex 4D scenes composed of moving Gaussians can be recovered by leveraging depth, segmentation, and tracking priors from 2D expert models [55, 51, 21]. Despite strong performance, these methods involve complex processing, including manual annotation on dynamic regions. It further requires lengthy optimization, limiting their scalability in practical applications.

Large reconstruction models. Transformer-based 3D large reconstruction models (LRMs) have shown strong potential for learning high-quality 3D reconstruction from data, at the object level [14, 16, 35, 58, 24] and static scenes [57, 67, 70, 59]. LRMs can generate reconstructions in seconds from a few input views, achieving quality comparable to optimization-based neural methods. However, training LRMs requires large-scale multi-view supervision of the same instance, which is readily available in synthetic datasets or static scene captures, but remains scarce for real-world videos.

Three recent efforts explored training transformers to predict time-dependent 3D-GS using animated synthetic data [42], self-curated real-world internet videos [25] and street-level data [61], making them the closest related works in motivation. In contrast to these methods, which predict time-dependent 3D-GS representations, our 4DGT offers a holistic 4D scene representation that captures geometry better and enables motion understanding capabilities lacking in prior approaches. [61] requires multi-camera input and only focuses on street-level scenes. Compared to training on synthetic data in [42], our real-world training approach generalizes better to real-world scenes.

3 Method

Given a posed monocular video, 4DGT uses a transformer to predict a set of 4DGS, which can be rendered in real time. We first describe the architecture and dynamic scene representation in section 3.1. To enable efficient training and rendering at scale, we introduce a pixel density control strategy and a level-of-detail structure based on spatial attention. Both techniques improve space-time sampling rates under fixed compute budgets. In section 3.3, we detail our training process and regularization strategies designed to resolve space-time ambiguities in monocular videos.

3.1 Feed-Forward Dynamic Gaussian Prediction

Input encoding. Given a posed monocular video, we extract a set of image frames \mathbf{I}_i with camera calibration \mathbf{P}_i and timestamp \mathbf{T}_i , denoted as $\{\mathbf{I}_i \in R^{H \times W \times 3}, \mathbf{P}_i \in R^{H \times W \times 6}, \mathbf{T}_i \in R^{H \times W \times 1} | i = 1 \dots N\}$, where \mathbf{P}_i represents the Plücker coordinates [18] and N is the total number of frames. We convert them into patches. For frame i , the patches are denoted as $\{\mathbf{I}_{i,j} \in R^{p \times p \times 3} | j = 1 \dots HW/p^2\}, \{\mathbf{T}_{i,j}\}$ and $\{\mathbf{P}_{i,j}\}$, where p is the patch size.

Feature fusion. We use the pretrained DINOv2 image encoder [33] to extract high-level C -dimensional features $\mathbf{F}_{i,j} \in R^C$. These are concatenated with the temporal and spatial encoding $\mathbf{T}_{i,j}$ and $\mathbf{P}_{i,j}$ as well as the input RGB image $\mathbf{I}_{i,j}$ to form the fused transformer input:

$$\{\mathbf{X}_{i,j}\} = \mathcal{F}(\{\mathbf{I}_{i,j} \oplus \mathbf{T}_{i,j} \oplus \mathbf{P}_{i,j} \oplus \mathbf{F}_{i,j} | i = 1 \dots N, j = 1 \dots HW/p^2\}), \quad (1)$$

where \oplus denotes the concatenation operation and \mathcal{F} denotes the all-to-all self-attention transformer module. In contrast to ViT, input used static LRMs in that it uses only Plücker rays [67] or DINO

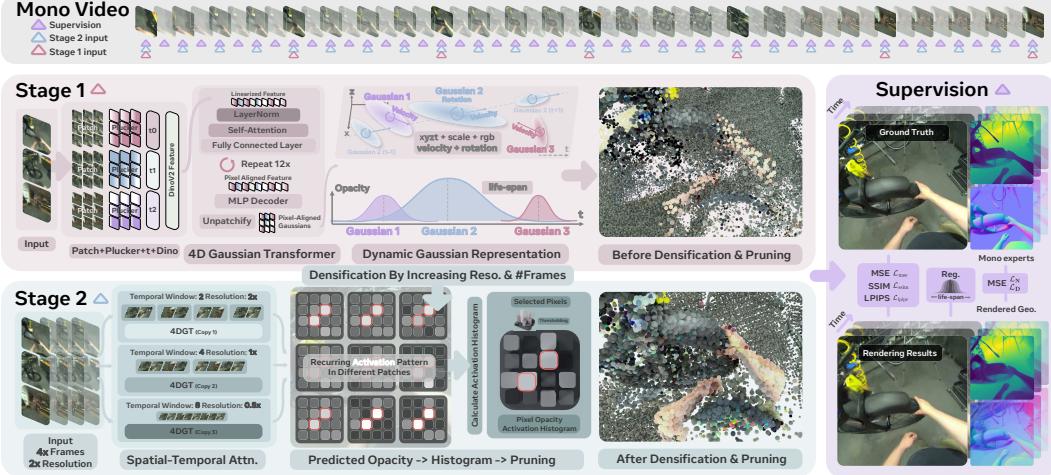


Figure 2: An overview of our method in training and rendering. 4DGT takes a series of monocular frames with poses as input. During training, we subsample the temporal frames at different granularity and use all images in training. We first train 4DGT to predict pixel-aligned Gaussians at coarse resolution in stage one. In stage two training, we pruned a majority of non-activated Gaussians according to the histograms of per-patch activation channels, and densify the Gaussian prediction by increasing the input token samples in both space and time. At inference time, we run the 4DGT network trained after stage two. It can support dense video frames input at high resolution.

feature [14], our transformer takes timestamp-aware Plücker rays with DINO feature together as input, which we found to be beneficial to provide the best prediction in view synthesis as well as geometry prediction.

Dynamic Gaussians. To represent geometry, we adopt the 2DGS [15] defined by the center $\mathbf{x} \in R^3$, scale $\mathbf{s} \in R^2$, opacity $\mathbf{o} \in R^1$ and orientation $\mathbf{q} \in R^4$ (quaternion) of the Gaussians. Compared to 3DGS[19] used in previous work [63, 53], 2DGS yields better geometry predictions. To represent motion, we use four temporal attributes, namely the temporal center $\mathbf{c} \in R^1$, life-span $\mathbf{l} \in R^1$, velocity $\mathbf{v} \in R^3$, and angular velocity $\omega \in R^3$ (as angle-axis) for each Gaussian. Given a specific timestamp t_s for rendering a particular dynamic Gaussian point $\mathbf{g} = \{\mathbf{x}, \mathbf{s}, \mathbf{q}, \mathbf{o}, \mathbf{c}, \mathbf{l}, \mathbf{v}, \omega\}$, we first calculate the offset to the opacity, location and orientation of the Gaussian point from the temporal attributes [63]. Specifically, the life-span \mathbf{l} is used to influence the Gaussian opacity \mathbf{o} over time:

$$\sigma = \sqrt{-\frac{1}{2} \cdot \frac{(l/2)^2}{\log(o_{th})}}, \quad o_{t_s} = \mathbf{o} \cdot e^{-\frac{1}{2} \cdot \frac{(t_s - c)^2}{\sigma^2}}, \quad (2)$$

where o_{th} is the opacity multiplier at the life-span boundary and σ is the standard deviation of the Gaussian distribution in the temporal domain. Intuitively, the Gaussian retains its full opacity at its temporal center and fades in a Gaussian distribution along the temporal axis. At $1/2$ time relative to the temporal center \mathbf{c} , the opacity of the point is reduced by multiplying a small factor o_{th} , which is set to 0.05 for all experiments. The location and orientation of each Gaussian is adjusted by the velocity \mathbf{v} and angular velocity ω to account for the motion:

$$\mathbf{x}_{t_s} = \mathbf{x} + \mathbf{v} \cdot (t_s - \mathbf{c}), \quad \mathbf{q}_{t_s} = \mathbf{q} \cdot \phi(\omega \cdot (t_s - \mathbf{c})), \quad (3)$$

where ϕ denotes converting the angle-axis representation to a quaternion. For each 2DGS with $\mathbf{x}_{t_s}, \mathbf{s}_{t_s}, \mathbf{q}_{t_s}, \mathbf{o}_{t_s}$ at timestamp t_s , we apply 2DGS rasterizer implemented in [65] to render image.

Dynamic Gaussian decoding. Given the pixel-aligned features $\{\mathbf{X}_{i,j}\}$, we use a transformer to decode the pixel-aligned 4DGS for each frame as

$$\{\mathbf{G}_{i,j}\} = \mathcal{D}_{\mathbf{x}, \mathbf{s}, \mathbf{q}, \mathbf{o}, \mathbf{c}, \mathbf{l}, \mathbf{v}, \omega}(\{\mathbf{X}_{i,j} | i = 1 \dots N, j = 1 \dots HW/p^2\}), \quad (4)$$

where $\mathcal{D}_{\mathbf{x}, \mathbf{s}, \mathbf{q}, \mathbf{o}, \mathbf{c}, \mathbf{l}, \mathbf{v}, \omega}$ denotes the MLP-based decoder head for producing the full suite of dynamic Gaussian parameters $\{\mathbf{G}_{i,j}\}$.

Our proposed 4DGS can unify the prediction of appearance and geometry properties of both static and dynamic elements. For static scenes, the network can learn to predict Gaussians with a long-living lifespan $1 \rightarrow \infty$, $\mathbf{v} \rightarrow 0$, $\omega \rightarrow 0$. For complex dynamic motions with occlusions, it predicts transient dynamic objects with short-living Gaussians $1 \rightarrow 0$.

3.2 Multi-level Pixel & Token Density Control

While pixel-aligned Gaussian has been a standard choice in prior work [67, 70], it has a severe limitation in representing video frames that require dense, long-term sampling to capture motion effectively. Naively sampling frames spatial-temporally would result in two key issues. First, the increasing number of aligned Gaussians degrades optimization and rendering performance, leading to suboptimal training and blurry details in dynamic regions. Second, the growing number of input tokens significantly increases computational cost, resulting in under-trained models.

Two-stage training. We introduce a two-stage approach to address these challenges. First, we train on coarsely sampled low-resolution images from scratch until convergence. In the second stage, inspired by 3DGS [19], we propose to filter pixel-aligned Gaussians by pruning low-opacity predictions per patch based on the histograms and increasing the token count to predict more Gaussians across space-time. Additionally, we introduce a multi-level spatiotemporal attention mechanism to further reduce the computational cost of self-attention layers.

Pruning. After the initial training stage at coarse resolution, we compute a histogram of activated Gaussians per patch and observe that only a few channels are activated. A similar pattern emerges in other pixel-aligned Gaussian methods [67], motivating us to decode only a small set of Gaussians using the activated channels. Formally, for each patch of Gaussian parameters $\mathbf{G}_{i,j}$, we consider the standard deviation of their opacity values $\mathbf{o}_{i,j} = \{\mathbf{o}_{i,j,k} | k = 1 \cdots p^2\}$:

$$\mu(\mathbf{o}_{i,j}) = \frac{1}{p^2} \sum_{k=1}^{p^2} \mathbf{o}_{i,j,k}, \quad \sigma(\mathbf{o}_{i,j}) = \sqrt{\mu(\mathbf{o}_{i,j}^2) - \mu(\mathbf{o}_{i,j})^2}, \quad (5)$$

where $\mu(\mathbf{o}_{i,j})$ is the mean of the opacity values and $\sigma(\mathbf{o}_{i,j})$ is the standard deviation. A particular pixel k is considered activated if it has a value larger than 1 unit of the standard deviation:

$$\mathbf{m}_{i,j,k} = \begin{cases} 1, & \mathbf{o}_{i,j,k} > \mu(\mathbf{o}_{i,j}) + \sigma(\mathbf{o}_{i,j}), \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

where k is the index of the pixel in the patch and $\mathbf{m}_{i,j,k}$ indicates whether the pixel is activated. A histogram $\mathbf{h}_{i,j}$ of all activation mask $\mathbf{m}_{i,j} = \{\mathbf{m}_{i,j,k} | k = 1 \cdots p^2\}$ for the patch output is:

$$\mathbf{h}_{i,j} = \sum_{i=1, j=1}^{N, HW/p^2} \mathbf{m}_{i,j}. \quad (7)$$

We select S channel from the histogram $\mathbf{h}_{i,j}$ for the patch output for all patches onward in training. This effectively implements an S/p^2 times reduction in the number of Gaussians for each patch, mimicking the pruning strategy of 3DGS [19]. We provide more in-depth analysis for this histogram-based pruning strategy compared to alternatives with visualizations in the supplementary materials.

Densification. The predicted Gaussian number can naturally increase with more space-time token inputs, either in resolution per frame or temporal frame numbers. The initially trained model provides a good scaffolding for pixel-aligned Gaussians when we increase the input token number in space and time. In the second stage of training, we increase the spatial and temporal resolution by a factor of R_s and R_t respectively. Combining the densification process and pruning strategy, this would result in $R_s^2 \cdot R_t \cdot S/p^2$ times of the Gaussians compared to the first stage. We select $R_s = 2$, $R_t = 4$, $S = 10$, and $p = 14$ for all experiments, leading to only 80% of the original number of Gaussians while greatly increasing the sampling rate of space-time by 16 times.

Multi-level spatial-temporal attention. The number of patches participating in the self-attention module \mathcal{F} increases by a factor of $R_s^2 \cdot R_t$ which will slow down optimization and inference significantly. To mitigate this, we propose a temporal level-of-detail attention mechanism to reduce the computational cost. We propose to divide the N input frames into M equal trunks in the highest level. This division limits the attention mechanism in the temporal dimension, but reduces

the computation of calculating n total tokens to $O(\frac{n^2}{M})$. To balance spatial-temporal samples, we construct a temporal level-of-detail structure by alternating the temporal range and spatial resolution, achieving a much smaller overhead while maintaining the ability to handle long temporal windows. For each level l , we reduce the spatial resolution by a factor of 2^l and increase temporal samples by 2. Empirically, we use level $L = 3$ and $M = 4$, which leads to an approximately 2 times reduction in the computational cost.

3.3 Training

Loss and regularization. We train 4DGT using segments of $W = 128$ consecutive frames from the monocular video and subsample every 8 frames as input, resulting in $N = 16$ input frames. Notably, for the second stage training where we apply techniques mentioned in section 3.2 and section 3.2, we increase the number of input frames to $N = 64$. After obtaining all Gaussian parameters $\{\mathbf{G}_{i,j}\}$ from each of the N input frames, we render them to all $W = 128$ images for self-supervision and compute the MSE loss. Additionally, we add the perceptual LPIPS loss [17] \mathcal{L}_{lpips} and SSIM loss [54] \mathcal{L}_{ssim} for better perceptual quality.

$$\mathcal{L}_{mse} = \sum_{i=1}^W \frac{\|\mathbf{I}_i - \mathbf{I}'_i\|_2}{W}, \quad \mathcal{L}_{lpips} = \sum_{i=1}^W \frac{\|\psi(\mathbf{I}_i) - \psi(\mathbf{I}'_i)\|_1}{W}, \quad \mathcal{L}_{ssim} = \sum_{i=1}^W \frac{SSIM(\mathbf{I}_i, \mathbf{I}'_i)}{W}, \quad (8)$$

where \mathbf{I}_i denotes the input image and \mathbf{I}'_i is the rendered image, ψ is the pre-trained layers AlexNet [20] and SSIM is the SSIM function [54]. To better regularize the training, we encourage the points to be static and have a long life-span using:

$$\mathcal{L}_v = \sum_{i=1, j=1, k=1}^{N, HW/p^2, p^2} \frac{\|\mathbf{v}_{i,j,k}\|_1}{NHW}, \quad \mathcal{L}_\omega = \sum_{i=1, j=1, k=1}^{N, HW/p^2, p^2} \frac{\|\boldsymbol{\omega}_{i,j,k}\|_1}{NHW}, \quad \mathcal{L}_l = \sum_{i=1, j=1, k=1}^{N, HW/p^2, p^2} \frac{\|\mathbf{l}_{i,j,k}\|_1}{NHW}, \quad (9)$$

where $\mathbf{v}_{i,j,k}$ is the velocity of the Gaussian point, $\boldsymbol{\omega}_{i,j,k}$ is the angular velocity of the Gaussian point and $\mathbf{l}_{i,j,k}$ is the life-span of the Gaussian point.

Expert guidance. We observe that training can benefit from leveraging monocular export models in geometry prediction. We extract the depth map \mathbf{D}_i and normal map \mathbf{N}_i from all W frames using DepthAnythingV2 [62] and StableNormal [64] and use them as a pseudo-supervision signal:

$$\mathcal{L}_D = \sum_{i=1}^W \frac{\|\mathbf{D}_i - \mathbf{D}'_i\|_2}{W}, \quad \mathcal{L}_N = \sum_{i=1}^W \frac{\|\mathbf{N}_i - \mathbf{N}'_i\|_2}{W}, \quad (10)$$

where \mathbf{D}'_i and \mathbf{N}'_i are the predicted depth and normal map rendered using the 2DGS rasterizer [15]. The final loss function for training the feed-forward prediction pipeline is:

$$\mathcal{L} = \mathcal{L}_{mse} + \lambda_{lpips} \mathcal{L}_{lpips} + \lambda_{ssim} \mathcal{L}_{ssim} + \lambda_v \mathcal{L}_v + \lambda_\omega \mathcal{L}_\omega + \lambda_l \mathcal{L}_l + \lambda_D \mathcal{L}_D + \lambda_N \mathcal{L}_N, \quad (11)$$

where λ_{lpips} , λ_{ssim} , λ_v , λ_ω , λ_l , λ_D and λ_N are the weights for the corresponding loss functions. We set $\lambda_{lpips} = 2.0$, $\lambda_{ssim} = 0.2$, $\lambda_v = 1.0$, $\lambda_\omega = 1.0$, $\lambda_l = 1.0$, $\lambda_D = 0.1$ and $\lambda_N = 0.01$ for all experiments. All weights for the regularization losses are warmed up linearly from 0 to their final values during the first 2500 iterations of training.

4 Implementation Detail

Architecture. We use a modified ViT architecture [7] for our fusion network \mathcal{F} . Specifically, we use 12 layers of all-to-all self-attention with 16 heads, each head having a hidden dimension of 96, and the fully connected layers have a $4 \times$ wider hidden channel size. Since the Plücker coordinates [18] \mathbf{P} and timestamps \mathbf{T} already provide the 4D position of each pixel, we do not use additional embedding for the positional information. For the second stage training, where we enable the multi-level spatial-temporal attention module, we copy the weights of the first-stage transformer L times and train them independently. The l -th transformer is responsible for the l -th level of spatial-temporal attention, with 1 classification token for passing information between different levels. For the MLP decoders \mathcal{D} , we use 2 fully connected layers with a hidden dimension of 256 for each channel. Both

the transformer modules \mathcal{F} and \mathcal{G} use GELU [13] as the activation function and layer normalization [1] as the normalization function. We also disable the bias parameters for all the layers.

Training & Inference. We implement 4DGT in PyTorch framework [38]. We employ FlashAttentionV3 [44] and the GSplat Rasterizer [65] for efficient attention and Gaussian optimization respectively. For optimization, we use the AdamW optimizer [27] with a learning rate of $5e^{-4}$ and a weight decay of 0.05. For the second stage training, the learning rate is set to $1e^{-5}$. Additionally, we linearly warm-up the learning rate of each stage in the first 2500 steps and then apply the cosine decaying schedule [26] for the remaining steps. During the second strange training, we additionally augment the input and output to the network by varying the aspect ratio and field of view of the images. Specifically, we randomly sample an aspect ratio from the uniform distribution on $[\frac{1}{3}, \frac{3}{1}]$ and a field of view ratio on the original image on [30%, 100%]. We train our reconstruction model 100k iterations for the first stage and 30k iterations for the second stage, using a total batch size of 64. With 64 Nvidia H100 GPUs, the first stage training takes roughly 9 days and the second stage training takes roughly 6 days. For all other experiments on inference speed, we use a single 80G A100 GPU.

5 Experiments

Training Datasets. We use the following real-world monocular videos with high-quality calibrations:

- Project Aria datasets with closed-loop trajectories: the EgoExo4D [12], Nymeria [29], Hot3D [2] and Aria Everyday Activities (AEA) [28].
- Video data with COLMAP [43] camera parameters: Epic-Fields [48, 6] and Cop3D [45].
- Phone videos with ARKit camera poses: ARKitTrack [69].

Evaluation datasets. We use the synthetic rendering provided in ADT [34] datasets, which provides metric ground truth depth. To evaluate cross-domain generalization, we use DyCheck [11] (DyC) datasets and the dynamic scene in TUM-SLAM [47] (TUM) to evaluate novel view synthesis. We further hold out a test split from EgoExo4D, AEA, and Hot3D, which we refer to as the Aria test set.

Metrics. For appearance evaluation, we compare the PSNR and LPIPS [68] metrics on novel view and time rendering results. For geometry evaluation, we compare against the depth RMSE [62] and normal angle error [64]. We additionally provide qualitative comparisons of motion rendered in 2D as optical flow and motion segmentation. All comparison experiments are conducted on 128-frame subsequences of the monocular videos, with 64 frames used as input and the remaining 64 frames used for testing, with images resized to 504×504 resolution or a similar pixel number for controlled comparison unless specified otherwise. For the DyCheck [11] dataset, we additionally compare the rendering results on the provided test view cameras, which shows signals on extreme view synthesis. We provide more details about evaluation implementations in the supplementary material.

Baselines. We consider the following baselines as most related work for evaluation.

1. **L4GM [42]:** It is the closest prior 4D Gaussian model that generalizes to real-world videos. Different from ours trained using real-world data only, they trained on a synthetic dataset and leveraged additional multi-view diffusion priors from ImageDream [50].
2. **Static-LRM:** We trained a static scene LRM following [67] on the same real world data as our 4DGT. We use 2DGS instead of 3DGS as the representation that resembles more similarity to our approach, except that we further model the dynamic content.
3. **Expert monocular models:** We compared each individual expert model we used during training in the same setting, including DepthAnythingV2 [62] aligned with the metric scale of UniDepth [39] and normals provided by StableNormal [64]. For novel view evaluation, we unproject the image using the nearest depth and normal frame.
4. **MonST3R [66]:** We compare to the dynamic point based representation [66] which highlights the representation difference in using 4DGS. We use ground truth camera poses as input to their model using the official implementation and using PyTorch3D [41] for normal estimation.
5. **Shape of Motion (SoM) [51]:** We use SoM to represent the top-tier per-scene optimization method as a reference for best dynamic reconstruction quality. We follow SoM’s instructions to manually segment the dynamic part. It requires running expert models as input, including mask, depth, and tracking, which we do not use. We include the preprocessing time for time comparisons.



Figure 3: From left-to-right, we show the novel space-time view comparisons on ADT [34], EgoExo4D [12], DyCheck [11] and the DyCheck test-view (rightmost). We render the depth (upper right) and normal (below right) next to each synthesized novel view. For ground truth depth and normal on EgoExo4D and DyCheck, we use predictions from the expert models from the ground truth image for reference. Please refer to supplementary materials for more visual comparisons.

Comparisons to baselines. Table 1 and Figure 3 present our comparisons to baselines. Compared to L4GM, which also predicts dynamic Gaussians from a trained transformer, 4DGT shows much better generalization across real scenes. We found that a static-LRM can provide a strong baseline for static scenes but will fail when dynamic motion is present, while 4DGT can do well in both. It is further validated in Table 2c on dynamic regions. The geometry predicted from 4DGT is more consistent with the world coordinate compared to expert models when evaluated in metric scale. Compared to optimization-based method SoM, 4DGT can offer on-par quality in view synthesis as well as geometry prediction while being 3 orders of magnitude faster in runtime, which makes it more favorable to process long-time videos in practice.

Ablation study in stage one training. In Table 2a for stage one training, we start from a *Naive* training baseline at coarse resolution using the image rendering losses in Eq. 8 trained only using AEA dataset with only 7 hours of data. After further scaling to using the EgoExo4D dataset (300 hours), we find that increasing the scale of the dataset can significantly improve the performance. We also compare to a static LRM [67] and per-frame LRM [42] counterpart in the same setting. We can already see the benefits over the baselines at this stage by a large margin. We further include the full training loss in Eq. 9 and Eq. 11, and we can see significant improvements in the quality across all metrics. Please refer to our supplementary for visual comparisons.

Table 1: Comparisons to baselines. We mark SoM in grey as reference for optimization based method, and rank the other baselines with ours as comparisons in learning based approaches.

Method	PSNR (Render) \uparrow				LPIPS (Render) \downarrow				RMSE (Depth) \downarrow			Deg. \downarrow	Recon. Time \downarrow
	ADT	TUM	DyC	Aria	Avg	ADT	TUM	DyC	Aria	Avg	ADT	TUM	Avg
SoM [51]	± 1.640	± 0.266	± 2.597	± 3.701		± 0.024	± 0.069	± 0.061	± 0.067		± 3.000	± 1.354	± 1.633
L4GM [42]	± 2.713	± 0.112	± 0.629	± 1.666		± 0.051	± 0.013	± 0.068	± 0.052		± 0.660	± 0.791	± 0.386
MonST3R [66, 64]	7.348	9.226	8.770	8.617	8.490	0.688	0.670	0.587	0.698	0.661	2.606	1.698	2.094
Experts [40, 64]	25.13	20.61	11.32	19.90	19.24	0.246	0.273	0.429	0.323	0.318	2.111	0.653	1.382
Ours	23.32	18.64	11.53	22.32	18.96	0.299	0.318	0.423	0.236	0.319	2.931	0.919	1.925
Ours	28.31	21.02	16.12	27.36	23.20	0.243	0.349	0.408	0.230	0.308	0.934	0.394	0.664

Table 2: Ablation study on our method components using ADT and DyCheck (DyC).

(a) Ablation on dynamic gaussian in stage one training. (b) Ablation on stage two training.

Method	PSNR \uparrow			LPIPS \downarrow			RMSE		Deg.		Method	PSNR \uparrow			LPIPS \downarrow			RMSE		Deg.	
	ADT	DyC	Avg	ADT	DyC	Avg	ADT	ADT \downarrow	ADT \downarrow	ADT \downarrow		Naive	ADT	DyC	Avg	ADT	DyC	Avg	ADT	ADT \downarrow	ADT \downarrow
Naive	15.49	13.95	14.72	0.612	0.517	0.564	1.156		42.25		Random.	25.79	14.97	20.38	0.333	0.480	0.406	0.750		25.84	
+ EgoExo4D [12]	22.72	15.27	19.00	0.229	0.385	0.307	1.278		41.11		+ D&P	28.79	15.52	22.16	0.242	0.434	0.338	0.722		25.95	
Static LRM [67]	19.29	14.21	16.75	0.399	0.463	0.431	0.830		31.96		+ Multi-level	28.71	15.34	22.03	0.242	0.439	0.341	0.783		27.30	
Per-frame [42]	10.07	12.10	11.08	0.748	0.714	0.731	3.117		61.77		+ Mix. (Ours)	28.31	16.12	22.22	0.243	0.408	0.326	0.934		25.92	
+ $\mathcal{L}_{N,D,l,\omega,v}$	26.45	15.86	21.15	0.170	0.399	0.284	0.773		21.59												

(c) Evaluation on the dynamic foreground.

Method	PSNR \uparrow			LPIPS \downarrow			RMSE		Deg.		Method	PSNR \uparrow			LPIPS \downarrow			RMSE		Deg.	
	ADT	DyC	Avg	ADT	DyC	Avg	ADT	ADT \downarrow	ADT \downarrow	ADT \downarrow		Naive	ADT	DyC	Avg	ADT	DyC	Avg	ADT	ADT \downarrow	ADT \downarrow
Static LRM [67] (masked)	17.30	13.56	16.76	0.059	0.220	0.102	0.613		49.35		Static LRM [67] (masked)	17.30	13.56	16.76	0.059	0.220	0.102	0.613		49.35	
Ours (masked)	27.29	14.93	22.86	0.030	0.195	0.075	0.388		33.32		Ours (masked)	27.29	14.93	22.86	0.030	0.195	0.075	0.388		33.32	

Ablation study in stage two training. Table. 2b shows the ablation study of key design in stage two training. Starting from a *Naive* model without density control to prune and densify (*D&P*) Gaussians and the *multi-level* attention proposed in 3.2, naively scaling up resolution and temporal samples will run out of memory in training. Compared with variants using a *Random* sampled Gaussian from predicted patches and further densifying, using the proposed *D&P* strategy to decode sparse activated Gaussian will lead to better results while being efficient in training. Adding the proposed multi-level attention can further speed up training with only a minor sacrifice in quality, but can speed up training two times faster. Finally, we mixed all the proposed datasets in training at stage two. Compared to model training only using EgoExo4D, mixing datasets improve generalization across domains.

6 Conclusion

We introduced 4DGT, a novel dynamic scene reconstruction method that feedforward predict 4DGS from an input posed video frame. The representation power of 4DGT enables it to handle general dynamic scenes using 4DGS with varying lifespans, and support it to handle complex dynamics in long videos. Different from prior work that heavily depends on multi-view supervision from synthetic datasets, 4DGT is trained only using real-world monocular videos. We demonstrate that 4DGT can generalize well to videos recorded from similar devices, and the ability of generalization can improve when mixing dataset training in scale.

Limitations and future work. We do not claim 4DGT can generalize to all videos in the wild. We assume the availability of a reliable calibration to train 4DGT and deploy it for inference. For this requirement, the training datasets have been limited to data sources from a few egocentric devices and phone captures. We observe that the quality may degrade in videos recorded by an unseen type of device due to the inaccurate metric scale calibrations. We believe this can be significantly improved by further scaling up the method using more diverse datasets recorded by different devices with curated calibrations. Similar to most monocular reconstruction methods, we still observe significant artifacts when viewing Gaussians from extreme view angles departing far from the input trajectory. Future directions can propose better representations to address it or learn to distill more priors from multi-view expert models, such as generative video models.

Contents

1	Introduction	2
2	Related Work	3
3	Method	3
3.1	Feed-Forward Dynamic Gaussian Prediction	3
3.2	Multi-level Pixel & Token Density Control	5
3.3	Training	6
4	Implementation Detail	6
5	Experiments	7
6	Conclusion	9
A	Additional Details	11
A.1	Additional Details on the Multi-Level Attention Module	11
A.2	Additional Details on the Densification & Pruning of the Dynamic Gaussian	11
A.3	Additional Details on Datasets and Baselines	12
B	Additional Results	12
B.1	Qualitative Results of the Ablation Study	12
B.2	Supplementary Videos	12
C	Social Impact	13

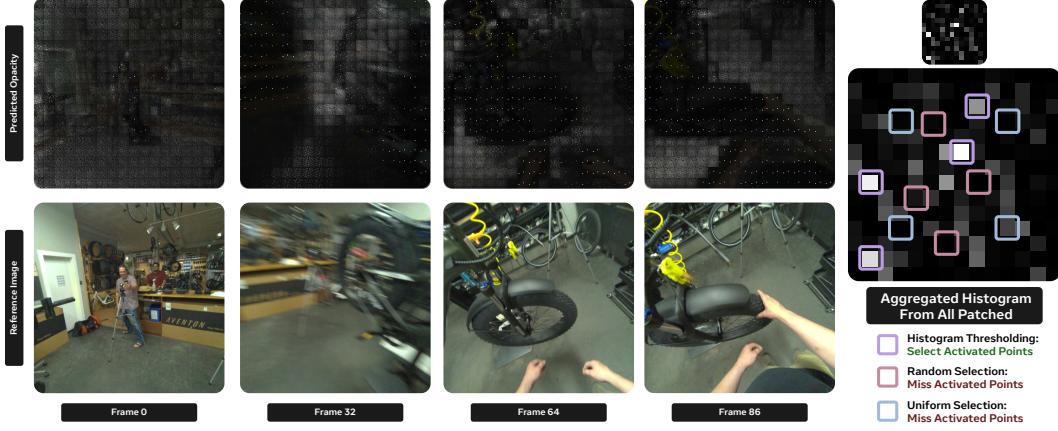


Figure 4: The predicted opacity map ($\in R^{N \times H \times W}$) of the pixel-aligned dynamic Gaussians from 4DGT and the computed histogram ($\in R^{p \times p}$) of the activation distribution. The right section shows the difference between histogram thresholding (Ours) and other filtering methods (randomly or uniformly selecting the Gaussians to keep) for reducing the number of Gaussians.

A Additional Details

A.1 Additional Details on the Multi-Level Attention Module

As mentioned in the Method section of the main paper, aside from the number of Gaussians, another efficiency-limiting factor is the number of tokens with the large number of high-resolution images. In the second stage of training, we increase the spatial and temporal resolution by a factor of R_s and R_t respectively. The number of patches participating in the self-attention module \mathcal{F} increases by a factor of $R_s^2 \cdot R_t$ which will slow down optimization and inference significantly. To mitigate this, we propose a temporal level-of-detail attention mechanism to reduce the computational cost. Specifically, noticing that the computational complexity of the self-attention module is $O(n^2)$ (simplified from $O(n^2 + n)$) where n is the number of tokens [49], We propose to divide the N input frames into M equal trunks in the highest level. This division limits the attention mechanism in the temporal dimension, but reduces the computation of calculating n total tokens to $O(\frac{n^2}{M})$. To balance spatial-temporal samples, we construct a temporal level-of-detail structure by alternating the temporal range and spatial resolution, achieving a much smaller overhead while maintaining the ability to handle long temporal windows. For each level l , we reduce the spatial resolution by a factor of 2^l and increase temporal samples by 2. This results in a computational complexity of:

$$O\left(\frac{n^2}{M} + \dots + \frac{n^2}{M \cdot 2^{L-1}}\right) = O\left(\frac{n^2}{M \cdot 2^{L-1}} \cdot \sum_{l=0}^{L-1} 2^l\right) = O\left(n^2 \cdot \frac{2^L - 1}{M \cdot 2^{L-1}}\right) \approx O\left(\frac{2n^2}{M}\right). \quad (12)$$

Empirically, we use level $L = 3$ and $M = 4$, which leads to an approximately 2 times reduction in the computational cost.

A.2 Additional Details on the Densification & Pruning of the Dynamic Gaussian

In appendix A.2, we visualize the predicted opacity map of the pixel-aligned dynamic Gaussians. It shows clear patterns of the activation of the pixel inside each patch, especially for the dynamic regions. Notably, randomly or uniformly selecting the Gaussians to keep will lead to a significant number of active Gaussians being pruned, effectively removing the ability of the model to model the dynamic parts, while our histogram thresholding scheme can effectively keep the Gaussians that are contributing. These strategies blend the densification and pruning strategies of Gaussian representations [19] and the multi-stage training strategy of ViT models [52], effectively introducing a density control scheme for the feed-forward prediction pipeline.

A.3 Additional Details on Datasets and Baselines

For each dataset used in training [28, 12, 2, 69, 29, 5, 48], we select 99.15% of the sequences as the training set and hold out the rest. For the datasets used in evaluation:

- **ADT** [34]: We select 4 subsequences for validating the reconstruction performance:
 - *Apartment_release_multiuser_cook_seq141_M1292*
 - *Apartment_release_multiskeleton_party_seq114_M1292*
 - *Apartment_release_meal_skeleton_seq135_M1292*
 - *Apartment_release_work_skeleton_seq137_M1292*
- **DyCheck** [11]: We use all 6 sequences with 3 views, and follow [51, 11] to apply the covisibility mask before computing metrics on novel views:
 - *apple, block, space-out, spin, paper-windmill, teddy*
- **TUM** [47]: We select 3 subsequences for evaluation:
 - *rgbd_dataset_freiburg2_desk_with_person*
 - *rgbd_dataset_freiburg3_walking_halfsphere*
 - *rgbd_dataset_freiburg3_sitting_halfsphere*
- **EgoExo4D** [12]: We select 3 subsequences from the hold-out sequences:
 - *cmu_bike01_2, sfu_cooking015_2, uniandes_bouldering_003_10*
- **Nymeria** [29]: We select 2 sequences from the hold-out set:
 - *20230607_s0_james_johnson_act1_7xwm28*
 - *20230612_s1_christina_jones_act0_u2r0z8*
- **AEA** [28]: We select the *loc5_script5_seq7_rec1* sequence from the hold-out set.
- **Hot3D** [2]: We select the *P0020_ff537251* sequence from the hold-out set.

The testing sequences from EgoExo4D, AEA, and Hot3D are denoted as *Aria* in all comparisons.

Note that we do not make comparisons with CAT4D [56], BulletTimer [25] and STORM [61] since they provide neither the source code nor the pre-trained models as of writing.

B Additional Results

B.1 Qualitative Results of the Ablation Study

In fig. 5, we show the ablation study results for the first and second stage training, respectively. As shown in the table and figure, our proposed loss and representation effectively model the dynamic regions and improve the reconstruction quality. Moreover, the proposed density control scheme effectively regularized the number of Gaussians with increased input count and input resolution, greatly improving details and avoiding using too much memory. Adding larger-scale datasets [12, 69] helps generalization for both in-domain and out-of-domain dataset.

B.2 Supplementary Videos

We attach additional video results in the supplementary video material. The supplementary video is structured as follows:

- 00:00:00-00:00:15: Brief introduction to the input & output setting and goal of the paper.
- 00:00:15-00:00:45: Reconstruction and novel view rendering results for the depth, normal, optical flow, dynamic mask, and appearance inferred in a rolling window fashion over a long video.
- 00:00:45-00:01:05: More qualitative video results from other datasets.
- 00:01:05-00:01:35: Comparison with baseline methods StaticLRM [67], L4GM [42] and Shape-of-Motion [51].
- 00:01:35-00:02:00: Ablation study of the proposed components.

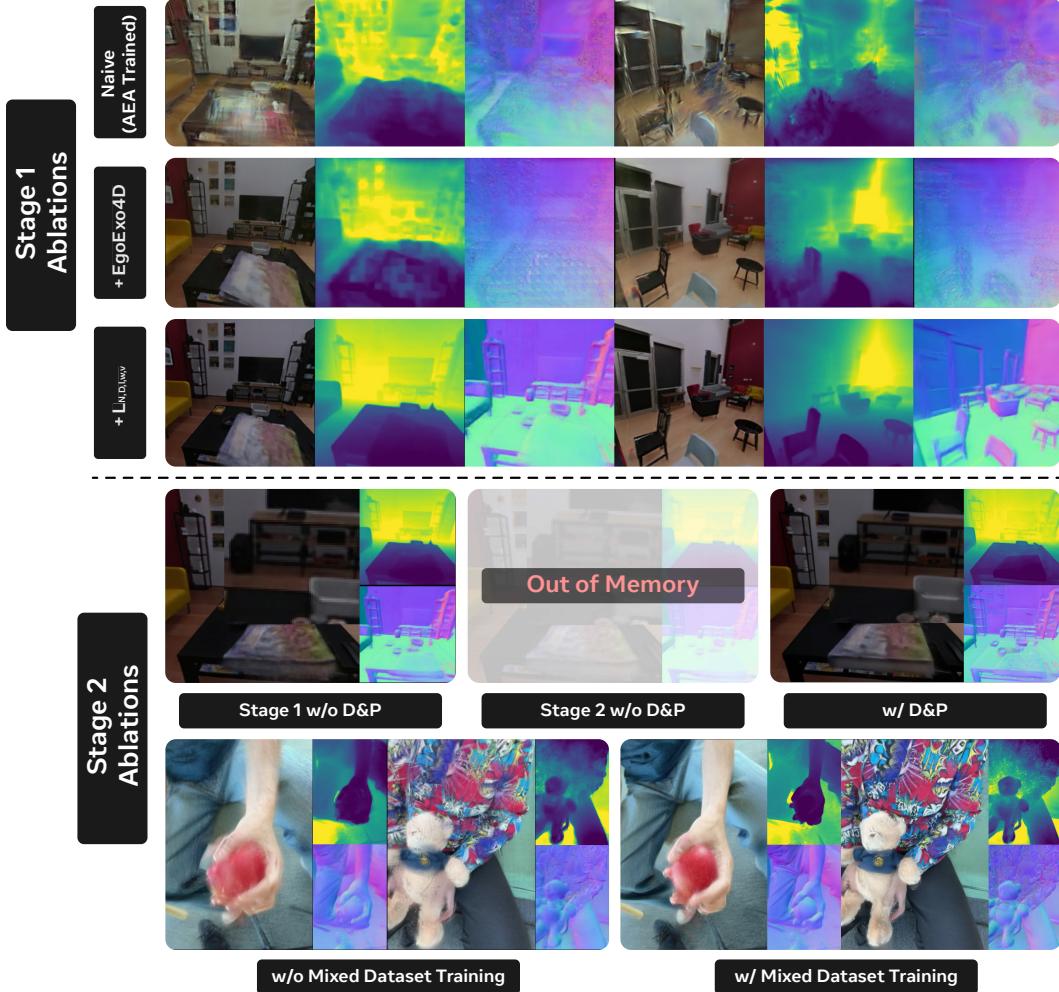


Figure 5: Ablation study on proposed components.

C Social Impact

As an early-stage research on 4D reconstruction, we do not foresee any immediate social impact from this work. However, it's worth noting that such a feed-forward pipeline could be used to synthesize more convincing fake videos by introducing novel views.

Acknowledgments and Disclosure of Funding

We thank Aljaz Bozic for the insightful discussions. We thank the Project Aria team for their open-source and dataset contributions.

References

- [1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [2] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. *arXiv preprint arXiv:2411.19167*, 2024.
- [3] A. Bozic, M. Zollhofer, C. Theobalt, and M. Niessner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

- [4] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)*, 39(4):86–1, 2020.
- [5] M. Caporuscio and A. Navarra. Cop3d: Context-aware overlay tree for content-based control systems. In *International Conference on Informatics in Control, Automation and Robotics*, volume 1, pages 305–310. SCITEPRESS, 2009.
- [6] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.
- [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Mininderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [8] Y. Duan, F. Wei, Q. Dai, Y. He, W. Chen, and B. Chen. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.
- [9] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Ghamo, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.
- [10] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.
- [11] H. Gao, R. Li, S. Tulsiani, B. Russell, and A. Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022.
- [12] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.
- [13] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- [14] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.
- [15] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024.
- [16] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao. Consistent4d: Consistent 360 { \backslash deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023.
- [17] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- [18] P. Julius. On a new geometry of space. *Julius Plückers Gesammelte wissenschaftliche Abhandlungen, op. cit.(24)*, 1:462–545, 1865.
- [19] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [21] J. Lei, Y. Weng, A. Harley, L. Guibas, and K. Daniilidis. MoScA: Dynamic gaussian fusion from casual videos via 4D motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024.
- [22] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022.
- [23] Z. Li, R. Tucker, F. Cole, Q. Wang, L. Jin, V. Ye, A. Kanazawa, A. Holynski, and N. Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. *arXiv preprint arXiv:2412.04463*, 2024.
- [24] Z. Li, D. Wang, K. Chen, Z. Lv, T. Nguyen-Phuoc, M. Lee, J.-B. Huang, L. Xiao, C. Zhang, Y. Zhu, et al. Lirm: Large inverse rendering model for progressive reconstruction of shape, materials and view-dependent radiance fields. *arXiv preprint arXiv:2504.20026*, 2025.
- [25] H. Liang, J. Ren, A. Mirzaei, A. Torralba, Z. Liu, I. Gilitschenski, S. Fidler, C. Oztireli, H. Ling, Z. Gojcic, et al. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. *arXiv preprint arXiv:2412.03526*, 2024.
- [26] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

- [27] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [28] Z. Lv, N. Charron, P. Moulon, A. Gamino, C. Peng, C. Sweeney, E. Miller, H. Tang, J. Meissner, J. Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.
- [29] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino, V. Baiyya, H. J. Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pages 445–465. Springer, 2024.
- [30] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [31] R. Murai, E. Dexheimer, and A. J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *CVPR*, June 2025.
- [32] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015.
- [33] M. Oquab, T. Dariseti, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [34] X. Pan, N. Charron, Y. Yang, S. Peters, T. Whelan, C. Kong, O. Parkhi, R. Newcombe, and Y. C. Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023.
- [35] Z. Pan, Z. Yang, X. Zhu, and L. Zhang. Efficient4d: Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742*, 2024.
- [36] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021.
- [37] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.
- [38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.
- [39] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024.
- [40] L. Piccinelli, C. Sakaridis, Y.-H. Yang, M. Segu, S. Li, W. Abbeloos, and L. Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025.
- [41] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.
- [42] J. Ren, C. Xie, A. Mirzaei, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim, H. Ling, et al. L4gm: Large 4d gaussian reconstruction model. *Advances in Neural Information Processing Systems*, 37:56828–56858, 2024.
- [43] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.
- [44] J. Shah, G. Bikshandi, Y. Zhang, V. Thakkar, P. Ramani, and T. Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.
- [45] S. Sinha, R. Shapovalov, J. Reizenstein, I. Rocco, N. Neverova, A. Vedaldi, and D. Novotny. Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. *CVPR*, 2023.
- [46] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *CVPR*, July 2017.
- [47] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgbd slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012.
- [48] V. Tschernezki, A. Darkhalil, Z. Zhu, D. Fouhey, I. Laina, D. Larlus, D. Damen, and A. Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *Advances in Neural Information Processing Systems*, 36:26485–26500, 2023.
- [49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

- [50] P. Wang and Y. Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.
- [51] Q. Wang, V. Ye, H. Gao, J. Austin, Z. Li, and A. Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024.
- [52] Q. Wang, Y. Zhang, A. Holynski, A. A. Efros, and A. Kanazawa. Continuous 3d perception model with persistent state, 2025.
- [53] Y. Wang, P. Yang, Z. Xu, J. Sun, Z. Zhang, C. Yuan, H. Bao, S. Peng, and X. Zhou. Freetimegs: Free gaussians at anytime anywhere for dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.
- [54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.
- [55] G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and W. Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023.
- [56] R. Wu, R. Gao, B. Poole, A. Trevithick, C. Zheng, J. T. Barron, and A. Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613*, 2024.
- [57] D. Xie, S. Bi, Z. Shu, K. Zhang, Z. Xu, Y. Zhou, S. Pirk, A. Kaufman, X. Sun, and H. Tan. Lrm-zero: Training large reconstruction models with synthesized data. *arXiv preprint arXiv:2406.09371*, 2024.
- [58] Y. Xie, C.-H. Yao, V. Voleti, H. Jiang, and V. Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024.
- [59] Y. Xu, Z. Shi, W. Yifan, H. Chen, C. Yang, S. Peng, Y. Shen, and G. Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024.
- [60] Z. Xu, Y. Xu, Z. Yu, S. Peng, J. Sun, H. Bao, and X. Zhou. Representing long volumetric video with temporal gaussian hierarchy. *ACM Transactions on Graphics*, 43(6), November 2024. URL <https://zju3dv.github.io/longvolcap>.
- [61] J. Yang, J. Huang, Y. Chen, Y. Wang, B. Li, Y. You, M. Igl, A. Sharma, P. Karkus, D. Xu, B. Ivanovic, Y. Wang, and M. Pavone. Storm: Spatio-temporal reconstruction model for large-scale outdoor scenes. *arXiv preprint arXiv:2501.00602*, 2025.
- [62] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.
- [63] Z. Yang, H. Yang, Z. Pan, and L. Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024.
- [64] C. Ye, L. Qiu, X. Gu, Q. Zuo, Y. Wu, Z. Dong, L. Bo, Y. Xiu, and X. Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 43(6):1–18, 2024.
- [65] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, J. Ye, J. Hu, M. Tancik, et al. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025.
- [66] J. Zhang, C. Herrmann, J. Hur, V. Jampani, T. Darrell, F. Cole, D. Sun, and M.-H. Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.
- [67] K. Zhang, S. Bi, H. Tan, Y. Xiangli, N. Zhao, K. Sunkavalli, and Z. Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *European Conference on Computer Vision*, 2024.
- [68] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. 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.
- [69] H. Zhao, J. Chen, L. Wang, and H. Lu. Arkittrack: A new diverse dataset for tracking using mobile rgb-d data. *CVPR*, 2023.
- [70] C. Ziwen, H. Tan, K. Zhang, S. Bi, F. Luan, Y. Hong, L. Fuxin, and Z. Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. *arXiv preprint arXiv:2410.12781*, 2024.