DynamicVerse: A Physically-Aware Multimodal Framework for 4D World Modeling

Kairun Wen[1*†], **Yuzhi Huang**[1*], **Runyu Chen**[1], **Hui Zheng**[1], **Yunlong Lin**[1], **Panwang Pan**[1], **Chenxin Li**[2], **Wenyan Cong**[3], **Jian Zhang**[1], **Junbin Lu**[4], **Chenguo Lin**[5], **Dilin Wang**[6], **Zhicheng Yan**[6], **Hongyu Xu**[6], **Justin Theiss**[6], **Yue Huang**[1], **Xinghao Ding**[1✉], **Rakesh Ranjan**[6], **Zhiwen Fan**[3]

* Equal Contribution; † Project Leader; ✉ Corresponding Author

[1]XMU [2]CUHK [3]UT Austin [4]UW [5]PKU [6]Meta

Project Website: `https://dynamic-verse.github.io/`

Figure 1: The overview of physically-aware multi-modal world modeling framework **DynamicVerse**.

Abstract

Understanding the dynamic physical world, characterized by its evolving 3D structure, real-world motion, and semantic content with textual descriptions, is crucial for human-agent interaction and enables embodied agents to perceive and act within real environments with human-like capabilities. However, existing datasets are often derived from limited simulators or utilize traditional Structure-from-Motion for up-to-scale annotation and offer limited descriptive captioning, which restricts the capacity of foundation models to accurately interpret real-world dynamics from monocular videos, commonly sourced from the internet.

To bridge these gaps, we introduce **DynamicVerse**, a physical-scale, multimodal 4D world modeling framework for dynamic real-world video. We employ large vision, geometric, and multimodal models to interpret metric-scale static geometry, real-world dynamic motion, instance-level masks, and holistic descriptive captions. By integrating window-based Bundle Adjustment with global optimization, our method converts long real-world video sequences into a comprehensive 4D multimodal format. DynamicVerse delivers a large-scale dataset consists of 100K+

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-3621

videos with 800K+ annotated masks and 10M+ frames from internet videos. Experimental evaluations on three benchmark tasks, namely video depth estimation, camera pose estimation, and camera intrinsics estimation, demonstrate that our 4D modeling achieves superior performance in capturing physical-scale measurements with greater global accuracy than existing methods.

1 Introduction

Humans inhabit a dynamic 3D world where geometric structure and semantic content evolve over time, constituting a 4D reality (spatial with temporal dimension). Understanding this dynamic environment is fundamental for developing advanced AI applications in fields such as robotics [1, 2, 3, 4, 5, 6], extended reality [7, 8, 9], and digital twins [10, 11]. However, building generalizable foundation models for these downstream tasks faces a longstanding challenge: acquiring high-quality, ground-truth 4D datasets from real-world environments, given that data-driven solutions increasingly demand 4D data while its collection using multiple sensors remains non-scalable. This raises the question: *Can we develop an automated pipeline capable of generating a real-world 4D dataset at scale?*

Current real-world 4D data primarily focus on indoor scenes [12, 13] or autonomous driving scenarios [14], where geometry capture is straightforward, but their diversity is limited. Even synthetic 4D data [15, 16, 17, 18], while controllable, often lack the fidelity and complexity required to truly represent the real world, resulting in a notable simulation-to-real gap. Moreover, physically-aware multimodal annotations—including metric-scale 3D geometry, detailed representations of non-rigid actors (*e.g.*, object size, mask and bounding box, etc.), and descriptive captions of dynamic contents (*i.e.*, object, camera and scene)—are often absent [19, 20]. This limited data landscape, especially when contrasted with the progress fueled by large-scale datasets in modalities like images, videos, and language, underscores *the compelling need for a large-scale, diverse, physically-aware, and semantically rich annotated multi-modal dataset for 4D scene understanding.*

Against this background, this paper aims to generate scalable, physically-aware, and multimodal annotations from massive monocular video data (see Fig. 1) for numerous potential applications, such as enhancing 4D Vision-Language Models [21], facilitating advanced 3D-aware video generation [22], and enabling linguistic 4D Gaussian Splatting [23]. However, achieving this goal is not trivial. To the best of our knowledge, there is currently a significant lack of rich and diverse 4D datasets (see Tab. 1) adequate for these demanding tasks. To address this data scarcity, we introduce **DynamicGen**, a novel automated data curation pipeline (see Fig. 3) designed to generate physically-aware multi-modal 4D data at scale. This pipeline contains two main stages: (1) metric-scale geometric and moving object recovery (*i.e.*, object category and mask) from raw videos, and (2) hierarchical dynamic contents (*i.e.*, object, camera and scene) detailed caption generation. Specifically, the pipeline curates diverse real-world monocular video sources; employs a filtering strategy to remove outliers such as camera motion intensity; integrates multiple foundation models (*i.e.*, VFMs, VLMs, LLMs, GFMs) for initial frame-wise annotation; applies dynamic bundle adjustment to jointly minimize global photometric error; and concludes with dynamic content captioning at three granularities and human-in-the-loop quality review to ensure annotation semantic accuracy.

The resulting multi-modal 4D dataset, termed **DynamicVerse** (see Fig. 1), comprises over 100K distinct 4D scenes, 800K masklets, and 10M video frames. Each scene is extensively annotated with multiple modalities: metric-scale point maps, camera parameters, object masks with corresponding categories, and detailed descriptive captions. We evaluate DynamicGen through three benchmarks: video depth estimation, camera pose estimation, and camera intrinsics estimation. We demonstrate the generalization capability of DynamicGen to process web-scale video data and extract multi-modal information qualitatively. We also conduct human study and GPT-assited evaluation to validate the quality of generated captions.

Our main contributions are summarized as follows:

- We develop **DynamicGen**, a novel automated data curation pipeline designed to generate physically-aware multi-modal 4D data at scale. This pipeline contains two main stages: (1) metric-scale geometric and moving object recovery from raw videos, and (2) hierarchical detailed semantic captions generation at three granularities (*i.e.*, object, camera and scene). Powered by foundation models (*i.e.*, VFMs, VLMs, LLMs, GFMs), DynamicGen efficiently generate 4D data

https://doi.org/10.52202/085713-3621

at scale, thus addressing the critical scalability, physical reality and modality diversity limitations of traditional 4D data curation.

- We introduce **DynamicVerse**, a large-scale 4D dataset featuring diverse dynamic scenes accompanied by rich multi-modal annotations including metric-scale point maps, camera parameters, object masks with corresponding categories, and detailed descriptive captions. DynamicVerse encompasses 100K+ 4D scenes coupled with 800K+ masklets, sourced through a combination of massive 2D video datasets and existing 4D datasets. This represents a significant improvement in terms of data scale, scene and modality diversity compared to prior 4D datasets.
- We validate DynamicGen through three benchmarks: video depth estimation, camera pose and intrinsics estimation. We demonstrate the generalization capability of DynamicGen to process web-scale videos and extract multi-modal information qualitatively. We also conduct human study and GPT-assited evaluation to validate the quality of generated captions.

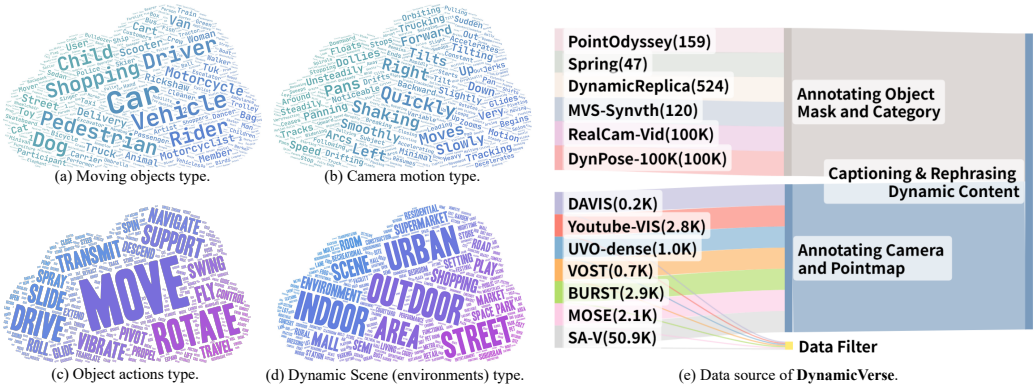
2 Related Work

Table 1: **Comparison of *DynamicVerse* with large-scale 2D video datasets and existing 4D scene datasets.** DynamicVerse expands the data scale and annotation richness compared to prior works.

	Numerical Statistics			*Provided Annotations*								*Detailed Features*			
Dataset Name	# Videos	# Frames	# Masklets	Camera	Depthmap	Instance Mask	Semantic Mask	Object Category	Object Caption	Scene Caption	Camera Caption	Scene Type	Dynamic Type	Real-world?	Metric-scale?
2D Video Dataset															
DAVIS2017 [24]	0.2K	10.7K	0.4K	✗	✗	✓	✓	✓	✗	✗	✗	-	-	-	-
Youtube-VIS [25]	3.8K	-	8,171	✗	✗	✓	✓	✓	✗	✗	✗	-	-	-	-
UVO-dense [26]	1.0K	68.3K	10.2K	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
VOST [27]	0.7K	75.5K	1.5K	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
BURST [28]	2.9K	195.7K	16.1K	✗	✗	✓	✓	✓	✗	✗	✗	-	-	-	-
MOSE [29]	2.1K	638.8K	5.2K	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
SA-V [30]	50.9K	4.2M	642.6K	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
MiraDATA [31]	330K	-	-	✗	✗	✓	✗	✓	✗	✗	✗	-	-	-	-
4D Scene Dataset															
T.Air Shibuya [32]	7	0.7K	-	✓	✓	✗	✓	✗	✗	✗	✗	Mixed	Street	Synthetic	Yes
MPI Sintel [33]	14	0.7K	-	✓	✗	✓	✗	✓	✗	✗	✗	-	Scripted	Synthetic	-
FlyingThings3D [34]	220	2K	-	✓	✗	✓	✓	✗	✗	✗	✗	Mixed	Objects	Synthetic	-
Waymo [14]	1,150	200K	-	✓	✓	✗	✗	✗	✗	✗	✗	Outdoor	Driving	Real-world	Yes
CoP3D [12]	4,200	600K	-	✓	✗	✓	✗	✗	✗	✗	✗	Mixed	Pets	Real-world	-
Stereo4D [35]	110,000	10,000K	-	✓	✓	✗	✗	✗	✗	✗	✗	Mixed	S. fisheye	Real-world	Yes
PointOdyssey [15]	159	200K	-	✓	✓	✓	✗	✗	✗	✗	✗	Mixed	Realistic	Synthetic	Yes
Spring [16]	47	6K	-	✓	✓	✓	✗	✗	✗	✗	✗	Mixed	Realistic	Synthetic	Yes
Dynamic Replica [17]	524	145K	-	✓	✓	✓	✗	✗	✗	✗	✗	Indoor	Realistic	Synthetic	Yes
MVS-Synth [18]	120	12K	-	✓	✓	✗	✗	✗	✗	✗	✗	Outdoor	Urban	Synthetic	Yes
RealCam-Vid [19]	100K	-	-	✓	✗	✓	✗	✗	✗	✓	✗	Mixed	Realistic	Synthetic	Yes
DynPose-100K [20]	100K	6,806K	-	✓	✗	✗	✗	✗	✗	✗	✗	Mixed	Realistic	Synthetic	Yes
DynamicVerse	100K+	13.6M	800K+	✓	✓	✓	✓	✓	✓	✓	✓	Mixed	Realistic	Real-world	Yes

Multi-modal foundation models. The development of numerous large foundation models in recent years has yielded remarkable performance across multiple tasks such as depth estimation [36, 37, 38, 39, 40], multi-view stereo [41, 42, 43], detection and segmentation [44, 45, 46, 30], human parsing [47], optical flow estimation [48, 49], and point tracking [50, 51, 38]. We propose that these models are highly applicable to achieving holistic 4D understanding, and unifying them within a single framework represents a promising direction for advancing tasks like nonrigid structure from motion. Our DynamicGen pipeline implements this idea by integrating the following pretrained components: UniDepthv2 [52] for geometry initialization, CoTracker3 [51] and UniMatch [49] for correspondence initialization, and Qwen2.5-VL [53] and SA2VA [54] for dynamic object segmentation. This integration, coupled with multi-stage optimization and regularization, allows us to extract accurate metric-scale camera poses and 4D geometry from monocular video. Similar to our method, the concurrently developed Uni4D [55] captures 4D geometry and pose, but it suffers from limited data modalities and discontinuous geometric estimates. In contrast, our *DynamicGen* pipeline not only produces globally refined dense 4D geometry but also supports moving object recovery (*i.e.*, object category and mask) and provides fine-grained dynamic content (*i.e.*, object, camera and scene) caption annotations.

https://doi.org/10.52202/085713-3621



(a) Moving objects type. (b) Camera motion type. (c) Object actions type. (d) Dynamic Scene (environments) type. (e) Data source of **DynamicVerse**.

Figure 2: The statistics and data source of **DynamicVerse**.

Multi-modal datasets. The development of large-scale multi-modal datasets has proven essential for advancing model performance across numerous domains, including language, image-text (*e.g.*, LAION [56, 57], Conceptual Captions [58], WebImageText [59]), and video understanding (*e.g.*, DAVIS2017 [24], Youtube-VIS [25], UVO-dense [26], VOST [27], BURST [28], MOSE [29], SA-V [30], MiraDATA [31]). Extending this success to holistic 4D understanding requires datasets that capture the dynamic 3D world with rich, multi-modal annotations. Existing 4D datasets, whether from early reconstruction efforts [15, 16, 17, 18] (limited diversity) or recent large-scale posed video collections like RealCam-Vid [19] and DynPose-100K [20] (lacking detailed geometry and semantics beyond pose), and even OBJAVERSE [60] (limited content), fall short of providing the comprehensive multi-modal information needed. Our *DynamicVerse* dataset bridges this gap by offering extensive multi-modal annotations, including metric-scale depth, camera parameters, instance segmentation with labels, and descriptive captions, specifically designed to facilitate advanced 4D research.

3 DynamicVerse

Overview DynamicVerse is a physical-scale, multi-modal 4D modeling framework for real-world video, which contains a novel automated data curation pipeline and corresponding large-scale 4D dataset. The **DynamicGen** pipeline (see Fig. 3) contains two main stages: (1) metric-scale geometric and moving object recovery (*i.e.*, object category and mask) from raw videos, and (2) hierarchical dynamic contents (*i.e.*, object, camera and scene) detailed caption generation. This pipeline primarily consists of five steps: 4D scene curation (in Sec. 3.1), data filter strategy (in Sec. 3.2), moving object recovery (in Sec. 3.3), dynamic bundle adjustment (in Sec. 3.4) and dynamic content caption generation (in Sec. 3.5). The resulting **DynamicVerse** dataset comprises over 100K distinct 4D scenes, 800K masklets, and 10M video frames. The data statistics and collection of DynamicVerse are illustrated in Fig. 2.

3.1 4D scene curation

To address the scarcity of available 4D scene data, DynamicGen unifies video data from various real-world video datasets, including DAVIS2017 [24], Youtube-VIS [25], UVO-dense [26], VOST [27], BURST [28], MOSE [29] and SA-V [30], alongside existing synthetic 4D datasets from PointOdyssey [15], Spring [16], Dynamic Replica [17], MVS-Synth [18], RealCam-Vid [19] and DynPose-100K [20]. The inclusion of these datasets is mainly motivated by their potential as scalable data sources for 4D scene understanding.

3.2 Data filter strategy

Data filtering is a critical step for identifying video data suitable for subsequent dynamic bundle adjustment. This process presents challenges due to the noisy quality and inherent variability of video data, which impedes the precise selection of high-quality sequences. To address this, we developed a filtering strategy incorporating several distinct criteria: proximal depth, focal-length stability, video blur, camera motion smoothness, and non-perspective distortion. Each of these aspects is quantified

Figure 3: The physically-aware multi-modal 4D data generation pipeline **DynamicGen**.

by a normalized score. We combine these scores as features and employ a Random Forest model to predict a video quality score ranging from 0 to 5. For model training, we manually annotated approximately 1,000 videos, assigning scores between 0 (indicating largely unsuitable, poor quality or insufficient dynamics) and 5 (indicating highly suitable, good quality and sufficient dynamics). We further apply VLM-based judgment to automatically exclude unsuitable videos before reconstruction.

3.3 Moving object recovery

To accurately identify the main dynamic objects within a video, we integrated multiple foundation models to achieve reliable segmentation. Specifically, our pipeline first employs Qwen2.5-VL [61] to identify moving objects and determine their semantic categories. These categories are then used to prompt SA2VA [54] for generating corresponding object masks. Leveraging the obtained object masks and geometric annotations, we can apply physical-aware size extraction to annotate the 3D bounding box for moving objects.

3.4 Dynamic bundle adjustment

Leveraging the high-quality RGB filtered videos, we employed a robust dynamic bundle adjustment method for annotating metric-scale camera parameters and point maps. This task is challenging due to dynamic objects occluding the static scene and static scene appearance changes hindering correspondence estimation. To effectively addresses both difficulties, we design a multi-stage optimization framework, see Fig. 3, including: (1) dynamic masking, (2) coarse camera initialization, (3) tracking-based static area bundle adjustment, (4) tracking-based non-rigid bundle adjustment, and (5) flow-based sliding window global refinement. Compared with traditional Structure-from-Motion techniques [62] and DUSt3R-based methods [63], our framework not only can handle massive video data with different resolutions but also yield metric-scale results by leveraging the full power of various foundation models.

Formulation Given T video RGB frames $\mathbf{I} = (I_1, \ldots, I_T)$ with resolution $H \times W$, we aim to estimate for each timestep $t = 1, \ldots, T$: per-frame pointmap $\mathbf{X}^t \in \mathbb{R}^{H \times W \times 3}$, camera intrinsics $\mathbf{K}^t$, and camera pose $\mathbf{P}^t = [\mathbf{R}^t | \mathbf{T}^t]$, where $\mathbf{R}^t$ and $\mathbf{T}^t$ denote the t-th camera's rotation and translation,

https://doi.org/10.52202/085713-3621

Figure 4: **Qualitative Results of Moving Object Segmentation.** We show qualitatively some of our segmentation results on the Youtube-VIS dataset compared with other methods.

respectively. Here, $\mathbf{X}$ contains static points $\mathbf{X}_{static}$ and dynamic points $\mathbf{X}_{dyn}$. We assume all frames share the same intrinsics K where we optimize focal lengths f_x and f_y. The overall cost function is formulated as follows:

$$C_{\text{BA}}(\mathbf{P}, \mathbf{X}_{\text{static}}) + C_{\text{flow}}(\mathbf{X}_{\text{static}}) + C_{\text{NR}}(\mathbf{X}_{\text{dyn}}) + C_{\text{motion}}(\mathbf{X}_{\text{dyn}}) + C_{\text{cam}}(\mathbf{R}) \tag{1}$$

where $C_{\text{BA}}(\mathbf{P}, \mathbf{X}_{\text{static}})$ and $C_{\text{flow}}(\mathbf{X}_{\text{static}})$ are bundle adjustment terms measuring the reprojection error between static correspondences and the static 3D structure $\mathbf{X}_{\text{static}}$. $C_{\text{NR}}(\mathbf{X}_{\text{dyn}})$ is a non-rigid structure-from-motion term evaluating the consistency of the dynamic point cloud with its tracklets. Regularization is applied to camera motion smoothness through $C_{\text{cam}}(\mathbf{P})$ and to the dynamic structure and motion via $C_{\text{motion}}(\mathbf{X}_{\text{dyn}})$. Each term participates in different optimization stages, which are described below. Detailed explanations of the cost terms are provided in the *supplementary material*.

Stage I: Dynamic masking We first extract dynamic masks to filter out the dynamic points for static area bundle adjustment. Specifically, we use semantic-based and motion-based method to obtain dynamic masks $\mathcal{M} = \{\mathbf{M}^t\}_{t=0}^T = \{\mathbf{M}_{sem}^t \cup \mathbf{M}_{flow}^t\}_{t=0}^T$. For the segmentation-based approach, we use the generated moving object masks $\{\mathbf{M}_{sem}^t\}_{t=0}^T$ in Sec. 3.3. For the flow-based approach, we employ Unimatch [49] to obtain dense optical flow predictions and compute per-frame epipolar error maps [64], which indicate the likelihood of pixels belonging to the dynamic foreground. Then we can obtain dynamic masks $\mathbf{M}_{flow} = [E_1, E_2, \ldots, E_T]$ by thresholding on these epipolar error maps.

https://doi.org/10.52202/085713-3621

Stage II: Coarse camera initialization In this stage, we start camera initialization by obtaining video depth $\mathcal{D} = \{\mathbf{D}_t\}_{t=0}^{T}$ and dense pixel motion $\mathcal{Z} = \{\mathbf{Z}_k\}_{k=0}^{K}$. For video depth estimation, we use UniDepthV2 [52], a monocular depth estimation network, to estimate initial depth maps $\mathcal{D}$ and initial camera intrinsics $\mathbf{K}_{\text{init}}$. For dense pixel motion estimation, we utilize Co-TrackerV3 [51] for its robustness. We apply Co-Tracker bi-directionally on a dense grid every 10 frames to ensure thorough coverage. We filter and classify tracklets using segmentation masks yielding a set of correspondent point trajectories $\{\mathbf{Z}_k \in \mathbb{R}^{T\times 2}\}_{k=0}^{K}$ at visible time steps determined by Co-Tracker. Combining $\mathcal{D}$ and $\mathcal{Z}$ allows us to establish 2D-to-3D correspondences. This allows us to initialize and tune camera parameter $\mathbf{P}$ by minimizing the following cost function with respect to camera parameters *only*. Specifically, we can unproject each video frame's depth at time t back to 3D and minimize the following cost function:

$$\min_{\mathbf{P}} \sum_{(t',t)} \sum_{\mathbf{Z}_k \in \neg\mathcal{M}} \|\mathbf{Z}_{k,t'} - \pi_{\mathbf{K}}(\pi_{\mathbf{K}}^{-1}(\mathbf{Z}_{k,t}, \mathbf{D}_t, \boldsymbol{\xi}_t), \boldsymbol{\xi}_{t'})\|_2^2 \tag{1}$$

where $\pi_{\mathbf{K}}^{-1}$ is the unprojection function that maps 2D coordinates into 3D world coordinates using estimated depth $\mathbf{D}_t$. We perform this over all pairs within a temporal sliding window of 5 frames. Given camera initialization $\hat{\mathbf{P}}$, we unproject our depth prediction into a common world coordinate system, which provides an initial 4D structure $\hat{\mathbf{X}}$. This is used as initialization for later optimization.

Stage III: Static area bundle adjustment We jointly optimizes camera pose and static geometry by minimizing the static component-related energy in a bundle adjustment fashion. Formally speaking, we solve the following:

$$\min_{\mathbf{P},\mathbf{X}_{\text{static}}} C_{\text{BA}}(\mathbf{P}, \mathbf{X}_{\text{static}}; \mathcal{Z}, \mathcal{M}) + C_{\text{cam}}(\mathbf{R}) \tag{2}$$

By enforcing consistency with each other, this improves both the static geometry and the camera pose quality. We perform a final scene integration by unprojecting correspondences into 3D using improved pose and filtering outlier noisy points in 3D.

Stage IV: Non-rigid bundle adjustment Given the estimated camera pose, this stage focuses on inferring dynamic structure. Note that we freeze camera parameters in this stage, as we find that incorrect geometry and motion evidence often harm camera pose estimation rather than improve it. Additionally, enabling camera pose optimization introduces extra flexibility in this ill-posed problem, harming robustness. Formally speaking, we solve the following:

$$\min_{\mathbf{X}_{\text{dyn}}} C_{\text{NR}}(\mathbf{X}_{\text{dyn}}; \mathbf{P}, \mathcal{Z}, \mathcal{M}) + C_{\text{motion}}(\mathbf{X}_{\text{dyn}}) \tag{3}$$

We initialize $\mathbf{X}_{\text{dyn}}$ using video depth and our optimized camera pose from last step. This energy optimization might still leave some high-energy noisy points, often from incorrect cues, motion boundaries, or occlusions. We filter these outliers based on their energy values in a final step. To further densify the global point cloud, enabling each pixel to correspond to a 3D point, we perform depthbased interpolation by computing a scale offset.

Stage V: Sliding window global refinement Given the estimated optical flow, this stage focuses on refining static structure. Note that we freeze camera parameters in this stage. Formally speaking, we solve the following:

$$\min_{\mathbf{X}_{\text{static}}} C_{\text{flow}}(\mathbf{X}_{\text{static}}) \tag{4}$$

With consideration for accuracy and efficiency, the sliding window global refinement is capable of significantly enhancing the multi-view consistency of static points and generalizing effectively to real-world 4D scenes. The detailed process can be found in the *appendix*.

3.5 Dynamic Content Caption Generation

Drawing upon the emphasis placed by LEO [65] and SceneVerse [66] on the criticality of caption quality and granularity for comprehensive scene understanding, we design captions at three specific levels: object, scene, and camera. Object captioning focuses on detailed object motion, scene captioning describes object-scene interactions, and camera captioning conveys intricate camera movement. To argument the caption, Large Language Models (LLMs) are employed to automatically rephrase initial captions and align them with these three granularity levels. Finally, to ensure data quality, human verification is conducted to filter out low-quality caption annotations.

 https://doi.org/10.52202/085713-3621

Moving object captioning. Moving object captions provide detailed descriptions crucial for object grounding. However, prior datasets often have incorrect temporal alignment [66] or insufficient detail [15, 67], while current video captioning methods yield only simple (*e.g.*, Panda-70M [68]) or non-localized descriptions (*e.g.*, Qwen2.5-VL [61]). To address these limitations and generate detailed, accurate captions for individual objects, we utilize DAM [69], known for its superior capabilities. Given RGB videos and corresponding object masks, DAM [69] generates detailed and temporally aligned object descriptions through carefully designed prompts, enabling precise grounding and richer scene understanding.

Dynamic scene captioning. Scene-level captions are designed to capture global information, depicting the key objects within the scene along with their associated actions, interactions, and functionalities. For a comprehensive understanding of the entire dynamic scene, we utilize Qwen2.5-VL [61] for dynamic scene captioning. To obtain more detailed, fine-grained, and accurate captions, we propose the use of structured captions. This process involves leveraging the fine-grained moving object captions as auxiliary input and employing specific prompting to generate the final scene-level descriptions. In the design of the prompts, we discovered that an explicit Hierarchical Prompt Design [70] significantly aids the Qwen2.5-VL[61] in comprehending its role, its expected format, and its operational boundaries. This approach contributes to the stabilization of the output's format and enhances the overall quality of the results.

Camera motion captioning. Camera Motion Captioning aims to describe the camera's trajectory and movement patterns. Using the powerful VLM [71], we analyze the sequence of inter-frame transformations to identify key motion types like panning, tilting, zooming, and dolly movements. This kinematic information is then used to generate natural language descriptions, potentially leveraging template-based generation or LLM prompting, to convey how the viewpoint changes over time.

Caption rephrasing. Following the generation of three distinct caption types (object, scene, and camera motion), a Large Language Model (LLM) [61] is employed to jointly process them. This step aligns the dynamic content descriptions across caption types and refines their phrasing to enhance overall consistency and readability.

Human-in-the-loop quality review. To provide a faithful comparison against larger pretrained models, human evaluation was used. Addressing persistent errors from source annotation inaccuracies, we implemented an iterative human-in-the-loop verification during caption construction to identify errors, trace sources, and revise/remove problematic data.

4 Experiments

In this section , we present experimental results to evaluate the robustness of our *DynamicGen* pipeline. Due to the page limit, we direct readers to the *appendix* for implementation details, more qualitative results, and more experimental analyses.

4.1 Video Depth Estimation

To evaluate video depth estimation accuracy, we assess several baseline methods, including metric depth predictors such as Metric3Dv2 [72], Depth-Pro [36], DepthCrafter [37], and Unidepth [39], which operate without scale or shift alignment. We also consider joint 4D modeling approaches, including MonST3R [63] and RCVD [73]. Evaluations are conducted on the Sintel [33] and KITTI [75] datasets, following standard protocols [37] by applying global shift and scale alignment to the predicted depth maps. We report absolute relative error (Abs Rel) and the percentage of inlier points ($\delta < 1.25$), with all methods undergoing least-squares alignment in disparity space. As shown in Tab. 2, *DynamicGen* achieves the best overall performance across all datasets and evaluation metrics. In particular, it consistently outperforms prior approaches in both absolute accuracy and geometric consistency, demonstrating strong generalization to diverse and dynamic scenes. As illustrated in Fig. 5, MonST3R consistently struggles with object geometry reconstruction, producing distorted

All research undertaken at Meta AI was limited to general guidance on model architectural design. Meta did not participate in any model training activities. Fan, Z. contributed to this project prior to the NeurIPS submission deadline.

https://doi.org/10.52202/085713-3621

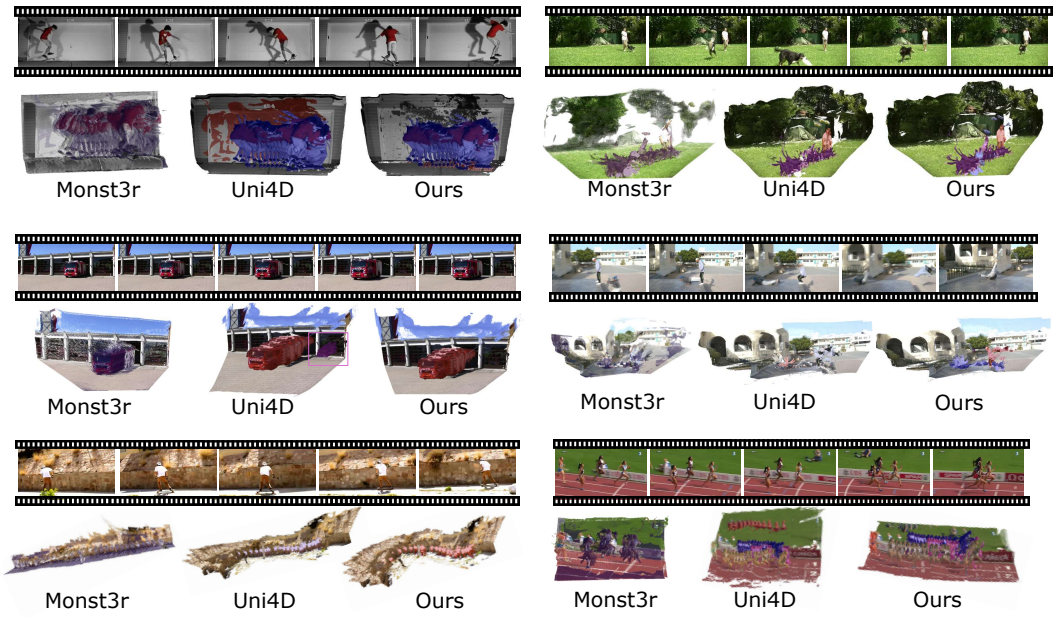


Figure 5: **Visual comparisons of 4D reconstruction on in-the-wild data.**

Table 2: Video depth evaluation on Sintel and KITTI datasets. **Bold** and underlined values indicate best and second best results.

			Sintel		KITTI	
Alignment	Category	Method	Abs↓	δ1.25↑	Abs↓	δ1.25↑
Per-sequence scale	Joint depth & pose	Monst3r [63]	0.344	55.9	0.089	91.4
		Uni4D [55]	**0.289**	**64.9**	**0.086**	**93.3**
Per-sequence scale & shift	Single-frame depth	Depth-pro [36]	0.280	60.5	0.080	94.2
		Metric3D [72]	**0.205**	**71.9**	**0.039**	**98.8**
	Video depth	DepthCrafter [37]	0.231	69.0	0.112	88.4
	Joint video depth & pose	Robust-CVD [73]	0.358	49.7	0.182	72.9
		CasualSAM [74]	0.292	56.9	0.113	88.3
		Uni4D [55]	0.216	72.5	0.098	89.7
		***DynamicGen*(Ours)**	**0.205**	**72.9**	**0.091**	**91.2**

shapes and noisy dynamic masks. Uni4D also exhibits mask imprecision. DynamicGen, however, achieves the cleanest dynamic segmentations and the strongest dynamic/static reconstructions.

4.2 Camera Pose Estimation

We evaluate our method against recent dynamic scene pose estimation approaches, including learning-based visual odometry (*e.g.*, LEAP-VO [76], DPVO [77]) and joint depth-pose optimization methods (*e.g.*, Robust-CVD [73], CasualSAM [74], MonST3R [63]). Experiments are conducted on the Sintel [33] and TUM-dynamics [78] datasets, following LEAP-VO's split for Sintel and subsampling the first 270 frames of TUM-dynamics, as done in MonST3R. Camera trajectories are aligned using Umeyama alignment [79], and we report Absolute Trajectory Error (ATE), Relative Translation Error (RPE trans), and Relative Rotation Error (RPE rot). As shown in Tab. 3, *DynamicGen* consistently achieves state-of-the-art results across all metrics and datasets, outperforming existing methods in both translation and rotation accuracy.

https://doi.org/10.52202/085713-3621

Table 3: Camera Pose Evaluation on Sintel and TUM-dynamic datasets. **Bold** and underlined values indicate best and second best results.

Category	Method	Sintel			TUM-dynamics		
		ATE↓	RPE trans↓	RPE rot↓	ATE↓	RPE trans↓	RPE rot↓
Pose only	DPVO [77]	0.171	**0.063**	**1.291**	**0.019**	**0.014**	**0.406**
	LEAP-VO [76]	**0.035**	0.065	1.669	0.025	0.031	2.843
Joint depth & pose	Robust-CVD [73]	0.368	0.153	3.462	0.096	0.027	2.590
	CasualSAM [74]	0.137	0.039	0.630	0.036	0.018	0.745
	Monst3r [63]	**0.108**	0.043	0.729	0.108	0.022	1.371
	Uni4D [55]	0.110	0.032	0.338	**0.012**	**0.004**	0.335
	***DynamicGen*(Ours)**	**0.108**	**0.029**	**0.282**	**0.012**	**0.004**	**0.331**

4.3 Camera Intrinsics Estimation

Camera intrinsics are typically unavailable for most casual videos, especially those sourced from the Internet. However, accurate intrinsics are critical for reliable pose estimation and 3D reconstruction. To assess this, we evaluate focal length estimation accuracy on the Sintel dataset, with results summarized in Tab. 4. UniDepth predicts depth and focal length from a single image, while Dust3r processes sequential frames but is trained under classical multi-view settings and fails to generalize well to dynamic scenes. In contrast, *DynamicGen* demonstrates strong generalization to dynamic content and achieves the best performance in both Absolute Focal Error (AFE) and Relative Focal Error (RFE), setting a new state-of-the-art for focal length estimation in unconstrained video scenarios.

Table 4: **Camera intrinsics estimation.**

Method	$AFE_{(px)}$↓	$RFE_{(\%)}$↓
UniDepth [39]	447.4	0.357
Dust3r [41]	434.0	0.364
***DynamicGen*(Ours)**	**413.1**	**0.241**

Table 5: **Dynamic Scene Caption evaluation.**

Method	Acc.↑	Com.↑	Con.↑	Rel.↑	Avg.↑
DO	79.28	76.65	73.23	80.33	77.37
+ SAKFE	80.23	77.46	74.01	81.45	78.29
+ HP	82.57	81.42	71.17	82.56	79.43
+ Rephrasing	82.48	80.50	71.86	83.27	79.53
+ COT	**84.38**	**82.09**	**75.87**	**85.56**	**81.97**

4.4 Caption Quality Evaluation

To assess caption quality, we sampled 100 videos from the SA-V dataset [30]. As presented in Table 5, our experimental results indicate that integrating semantic-aware key frame extraction (SAKFE), hierarchical prompting (HP), caption rephrasing, and Chain-of-Thought (CoT) prompting [80] significantly enhances the quality of dynamic scene captions generated by Vision-Language Models (VLMs). We evaluated caption quality using the LLM-as-Judge metric G-VEval [81], conducting ten independent evaluations to ensure robust average results. The resulting captions were demonstrably more accurate, complete, concise, and relevant than those produced by direct output (DO), confirming the effectiveness of these strategies for improving caption quality in this task.

5 Conclusion

In this work, we address key limitations in traditional 4D data curation regarding scalability, physical realism, and modality diversity. We introduce DynamicGen, an automated pipeline leveraging foundation models for video filtering, metric-scale geometry and motion recovery, and hierarchical semantic captioning from raw videos. DynamicGen's capabilities are validated through standard benchmarks on video depth and camera pose/intrinsics estimation, qualitative analyses on diverse web videos, and human/LLM-based evaluations confirming caption quality. Utilizing DynamicGen, we construct DynamicVerse, a large-scale 4D dataset with over 100K dynamic scenes and rich physically grounded multimodal annotations. Together, this work offers a scalable 4D data generation methodology and a comprehensive new resource to advance 4D scene understanding.

References

[1] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015.

[2] Chao Yu, Zuxin Liu, Xin-Jun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. Ds-slam: A semantic visual slam towards dynamic environments. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 1168–1174. IEEE, 2018.

[3] Berta Bescos, José M Fácil, Javier Civera, and José Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE robotics and automation letters*, 3(4):4076–4083, 2018.

[4] Linhui Xiao, Jinge Wang, Xiaosong Qiu, Zheng Rong, and Xudong Zou. Dynamic-slam: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. *Robotics and Autonomous Systems*, 117:1–16, 2019.

[5] Jiahui Huang, Sheng Yang, Zishuo Zhao, Yu-Kun Lai, and Shi-Min Hu. Clusterslam: A slam backend for simultaneous rigid body clustering and motion estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2019.

[6] Jesse Morris, Yiduo Wang, and Viorela Ila. The importance of coordinate frames in dynamic slam. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13755–13761. IEEE, 2024.

[7] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[9] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023.

[10] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. *Advances in Neural Information Processing Systems*, 37:74383–74410, 2024.

[11] Hezhen Hu, Zhiwen Fan, Tianhao Wu, Yihan Xi, Seoyoung Lee, Georgios Pavlakos, and Zhangyang Wang. Expressive gaussian human avatars from monocular rgb video. *arXiv preprint arXiv:2407.03204*, 2024.

[12] Samarth Sinha, Roman Shapovalov, Jeremy Reizenstein, Ignacio Rocco, Natalia Neverova, Andrea Vedaldi, and David Novotny. Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4881–4891, 2023.

[13] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram 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.

[14] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[15] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023.

 https://doi.org/10.52202/085713-3621

[16] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2023.

[17] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13229–13239, 2023.

[18] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2821–2830, 2018.

[19] Guangcong Zheng, Teng Li, Xianpan Zhou, and Xi Li. Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *arXiv preprint arXiv:2504.08212*, 2025.

[20] Chris Rockwell, Joseph Tung, Tsung-Yi Lin, Ming-Yu Liu, David F Fouhey, and Chen-Hsuan Lin. Dynamic camera poses and where to find them. *arXiv preprint arXiv:2504.17788*, 2025.

[21] Hanyu Zhou and Gim Hee Lee. Llava-4d: Embedding spatiotemporal prompt into lmms for 4d scene understanding, 2025.

[22] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025.

[23] Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, and Hanspeter Pfister. 4d langsplat: 4d language gaussian splatting via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[24] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[25] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5188–5197, 2019.

[26] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10776–10785, 2021.

[27] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the" object" in video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22836–22845, 2023.

[28] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1674–1683, 2023.

[29] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20224–20234, 2023.

[30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[31] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

[32] Yuheng Qiu, Chen Wang, Wenshan Wang, Mina Henein, and Sebastian Scherer. Airdos: Dynamic slam benefits from articulated objects. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8047–8053. IEEE, 2022.

[33] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012.

[34] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[35] Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4d: Learning how things move in 3d from internet stereo videos. *arXiv preprint arXiv:2412.09621*, 2024.

[36] Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

[37] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024.

[38] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[39] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher 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] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[41] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.

[42] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.

[43] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[44] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

[45] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[46] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

 https://doi.org/10.52202/085713-3621

[47] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024.

[48] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022.

[49] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[50] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024.

[51] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024.

[52] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025.

[53] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[54] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

[55] David Yifan Yao, Albert J Zhai, and Shenlong Wang. Uni4d: Unifying visual foundation models for 4d modeling from a single video. *arXiv preprint arXiv:2503.21761*, 2025.

[56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[57] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[58] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[59] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021.

[60] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.

[61] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,

Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[62] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European Conference on Computer Vision*, pages 523–542. Springer, 2022.

[63] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024.

[64] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023.

[65] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

[66] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024.

[67] Wanhua Li, Renping Zhou, Jiawei Zhou, Yingwei Song, Johannes Herter, Minghan Qin, Gao Huang, and Hanspeter Pfister. 4d langsplat: 4d language gaussian splatting via multimodal large language models. *arXiv preprint arXiv:2503.10437*, 2025.

[68] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.

[69] Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, and Yin Cui. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*, 2025.

[70] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

[71] Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video. *arXiv preprint arXiv:2504.15376*, 2025.

[72] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. 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*, 2024.

[73] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021.

[74] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022.

[75] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

 https://doi.org/10.52202/085713-3621

[76] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19844–19853, 2024.

[77] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36:39033–39051, 2023.

[78] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012.

[79] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991.

[80] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[81] Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7419–7427, 2025.

[82] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116. Citeseer, 2007.

[83] Wei-Chiu Ma, Shenlong Wang, Rui Hu, Yuwen Xiong, and Raquel Urtasun. Deep rigid instance scene flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3614–3622, 2019.

[84] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022.

[85] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024.

[86] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction sections offer a comprehensive discussion of the manuscript's context, intuition, and ambitions, as well as its contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed by authors at the end of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/085713-3621

Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The pipeline of the methods and the details of experiments are presented with corresponding reproducible credentials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All utilized data are sourced from open-access platforms. The code, which will be made publicly available, is uploaded as a zip file.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The pipeline of the methods and the details of experiments are presented with corresponding reproducible credentials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The results contain the standard deviation of the results over several random runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

 https://doi.org/10.52202/085713-3621

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details of experiments are presented with corresponding reproducible credentials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original owners of assets, including data and models, used in the paper, are properly credited and are the license and terms of use explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

https://doi.org/10.52202/085713-3621

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced in the paper are well documented and provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: as a controller

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-3621

Figure 6: DynamicVerse dataset.

A Appendix

In the appendix, we provide more results and analysis and summarize them as follows:

- In Section A.1, we introduce the broader impact of our DynamicVerse framework.
- In Section A.2, we supplement details of dynamic bundle adjustment.
- In Section A.3, we ablate the different components for dynamic bundle adjustment.
- In Section A.4, we provide additional experiments on generated hierarchical captions.
- In Section A.5, we provide more qualitative results of dynamic bundle adjustment.
- In Section A.6, we provide inference speed and computational cost for DynamicGen.
- In Section A.7, we provide the limitation.

A.1 Broader Impact

The introduction of DynamicVerse, with its large-scale, physically-aware, and multimodally annotated 4D dataset derived from real-world videos, is set to significantly influence several advanced research areas. Our framework's unique ability to capture metric-scale geometry, real-world motion, instance-level semantics, and descriptive captions offers an unparalleled resource that can catalyze progress in the following domains:

- **Dynamic 4D Scene Generation**: DynamicVerse offers a paradigm shift for Dynamic 4D Scene Generation. Current methods often rely on limited simulators or struggle to realistically portray complex real-world physics and motion from internet-sourced content. By accurately interpreting real-world dynamics from monocular videos and integrating window-based Bundle Adjustment with global optimization, DynamicVerse converts long video sequences into a comprehensive 4D multimodal format, capturing fine-grained dynamic information. This rich, real-world data provides an unparalleled training ground for generative models, leading to the creation of highly

realistic, physically plausible, and semantically coherent dynamic 4D scenes. This has profound implications for high-fidelity content creation in entertainment (e.g., movies, games), realistic virtual environments for training and simulation (e.g., disaster response, architectural visualization), and the synthetic generation of diverse data for further AI research, helping to overcome privacy and data collection limitations.

- **4D Vision-Language Models (4D-VLM)**: DynamicVerse will greatly accelerate the development of sophisticated 4D Vision-Language Models that can reason about space, time, and semantics concurrently. Existing VLMs often operate on 2D images or short video clips with limited 3D awareness. Our framework provides a unique combination of metric-scale 4D geometry, real-world dynamic motion, and comprehensive textual descriptions for long video sequences, allowing 4D-VLMs to learn intricate relationships between evolving 3D scenes and natural language narratives. Such models could enable more advanced human-agent interaction, where agents can provide detailed textual explanations of complex dynamic events they perceive in 4D, or understand nuanced, temporally extended instructions involving interactions within a 3D space. This could revolutionize areas like AI-powered video captioning, temporal question answering in 3D, and the development of embodied AI agents that communicate their understanding of the dynamic world with human-like richness.
- **4D Language-Grounded Gaussian Splatting (4D-LangSplat)**: DynamicVerse offers a foundational dataset for advancing 4D-LangSplat methodologies. While current 4D Gaussian Splatting techniques excel at novel view synthesis of dynamic scenes, their integration with language for semantic understanding and manipulation is still nascent. Our dataset, rich with 800K+ instance masks and holistic descriptive captions directly linked to evolving 3D structures and motions at a physical scale, empowers 4D-LangSplat models. This will enable the development of systems that can not only reconstruct dynamic scenes with high fidelity but also allow users to query, edit, and interact with these 4D representations using natural language. For instance, users could ask an agent to "track the red car that just turned left" or "remove the person walking in front of the fountain," with the model understanding both the spatial dynamics and the semantic context. This can significantly enhance applications in robotics, augmented reality, and interactive content creation by bridging the gap between visual perception and linguistic instruction in dynamic 3D environments.

In summary, DynamicVerse is poised to serve as a crucial catalyst, providing the data and framework necessary to bridge the gap between 2D understanding and true 4D world modeling, thereby fostering advancements in semantic scene understanding, dynamic object interaction, multimodal reasoning, and realistic content generation.

A.2 Details of Dynamic Bundle Adjustment

Camera parameterization In Eq. (1), $\boldsymbol{\xi} \in \mathbb{SE}(3)$ represents the camera poses as rigid transformations. Rotations are parameterized using $so(3)$ rotation vectors, which offer a minimal representation facilitating direct optimization.

Static Area Bundle Adjustment Term In Eq. (2), the bundle adjustment energy $C_{\mathrm{BA}}(\mathbf{P}, \mathbf{X}_{\mathrm{static}})$ measures the consistency between the pixel-level correspondences and the 3D structure of static scene elements. Given the input pixel tracks $\mathcal{Z} = \{\mathbf{Z}_k\}_{k=0}^{K}$ and video segmentation $\mathcal{M} = \{\mathbf{M}^t\}_{t=0}^{T}$, we filter all tracks corresponding to static areas and minimize the distance between the projected pixel location and the observed pixel location:

$$C_{\mathrm{BA}}(\mathbf{P}, \mathbf{X}_{\mathrm{static}}; \mathcal{Z}, \mathcal{M}) = \sum_{\mathbf{Z}_k \in \mathcal{M}} \sum_{t} w_{k,t} \|\mathbf{Z}_{k,t} - \pi_K(\mathbf{X}_k, \boldsymbol{\xi}_t)\|_2 \tag{5}$$

where $\mathbf{X}_k$ is the k-th 3D point, $\mathbf{Z}_{k,t}$ is the k-th 3D point's corresponding pixel track's 2D coordinates at time t, $w_{k,t} \in \{0, 1\}$ is a visibility indicator and π_K is the perspective projection function.

Camera Smoothness Prior In Eq. (2), given the video input, a temporal smoothness prior is imposed on camera poses. This prior penalizes abrupt changes in relative pose, defined as $\xi_{t \to t+1} = \xi_{t+1}^{-1} \cdot \xi_t$. We adaptively reweight this term based on the magnitude of the relative motion. Specifically, a larger relative motion results in a reduced penalty on its change rate, while a smaller relative motion

 https://doi.org/10.52202/085713-3621

incurs a higher penalty. Formally, this is expressed as:

$$C_{\text{cam}}(\mathbf{P}) = \sum_t C_{\text{rot}}(\mathbf{R}_{t-1,t,t+1}) + \sum_t C_{\text{trans}}(\mathbf{T}_{t-1,t,t+1})$$

where $C_{\text{rot}}(\mathbf{R}_{t-1,t,t+1}) = \frac{2||\text{rad}(\mathbf{R}_{t\to t+1}) - \text{rad}(\mathbf{R}_{t-1\to t})||}{||\text{rad}(\mathbf{R}_{t-1\to t})|| + ||\text{rad}(\mathbf{R}_{t\to t+1})||}$ and $C_{\text{trans}}(\mathbf{t}_{t-1,t,t+1}) = \frac{2||\mathbf{t}_{t\to t+1} - \mathbf{t}_{t-1\to t}||}{||\mathbf{t}_{t-1\to t}|| + ||\mathbf{t}_{t\to t+1}||}$; rad converts the rotation matrix into absolute radians.

Non-Rigid Bundle Adjustment Term In Eq. (3), for dynamic objects, we impose a nonrigid bundle adjustment term, $E_{\text{NR}}(\mathbf{X}_{\text{dyn}})$, which measures the discrepancy between the dynamic point cloud and pixel tracklets. Here, each pixel tracklet corresponds to a dynamic 3D point sequence, $\{\mathbf{X}_{k,t}\}$, optimized for each observed tracklet:

$$C_{\text{NR}}(\mathbf{X}_{\text{dyn}}, \mathbf{P}, \mathcal{Z}, \mathcal{M}) = \sum_{\mathbf{z}_k \in \mathcal{M}} \sum_t w_{k,t} \|\mathbf{Z}_{k,t} - \pi_K(\mathbf{X}_{k,t}, \boldsymbol{\xi}_t)\|_2 \tag{6}$$

where $\mathbf{X}_{k,t} \in \mathbb{R}^3$ is the k-th dynamic point's location at t.

Dynamic Motion Prior In Eq. (3), $C_{\text{motion}}(\mathbf{X}_{\text{dyn}})$ is a regularization term that encodes the characteristics of the dynamic structure. It contains two prior terms that are used to regularize the dynamic structure, both of which have demonstrated effectiveness in previous work.

$$C_{\text{motion}}(\mathbf{X}_{\text{dyn}}) = C_{\text{arap}}(\mathbf{X}_{\text{dyn}}) + C_{\text{smooth}}(\mathbf{X}_{\text{dyn}}). \tag{7}$$

C_{arap} represents an as-rigid-as-possible (ARAP) prior [82] designed to penalize extreme deformations that compromise local rigidity. Specifically, for each dynamic control point k, its nearest neighbors are identified using k-Nearest Neighbors (KNN) on the remaining tracks. We then enforce that the relative distances among these neighboring pairs remain consistent, preventing sudden changes

$$C_{\text{arap}} = \sum_t \sum_{(k,m)} w_{km} \|d(\mathbf{X}_{k,t}, \mathbf{X}_{m,t}) - d(\mathbf{X}_{k,t+1}, \mathbf{X}_{m,t+1})\|^2 \tag{8}$$

where $d(,)$ is the L2 distance and $w_{km,t} = 1$ if all relevant points are visible.

C_{smooth} is a simple smoothness term that promotes temporal smoothness for the dynamic point cloud:

$$C_{\text{smooth}} = \sum_{\mathbf{t}} \sum_{\mathbf{X}_k \in \mathbf{X}_{\text{dyn}}} w_{k,t} \|\mathbf{X}_{k,t} - \mathbf{X}_{k,t+1}\|_2. \tag{9}$$

Despite simplicity, both motion terms are crucial in our formulation, as they significantly reduce ambiguities in 4D dynamic structure estimation, which is highly ill-posed. Unlike other methods, we do not assume strong modelbased motion priors, such as rigid motion [83], articulated motion [84], or a linear motion basis [85].

Optical Flow Prior In Eq. (4), we also use a flow projection loss to encourage the global pointmaps to be consistent with the estimated flow for the confident, static regions of the actual frames. More precisely, given two frames t, t', using their global pointmaps, camera extrinsics and intrinsics, we compute the flow fields from taking the global pointmap $\mathbf{X}_{\text{t}}$, assuming the scene is static, and then moving the camera from t to t'. We denote this value $\mathbf{F}_{\text{cam}}^{\text{global: } t\to t'}$, similar to the term defined in the confident static region computation above. Then we can encourage this to be close to the estimated flow, $\mathbf{F}_{\text{est}}^{\text{t}\to\text{t}'}$, in the regions which are confidently static $\mathbf{X}_{\text{staic}}^{\text{global: } t\to t'}$ according to the global parameters:

$$C_{\text{flow}}(\mathbf{X}_{\text{static}}) = \sum_{W^i \in W} \sum_{t' \in W^i} \|\mathbf{X}^{\text{global: } t\to t'} \cdot (\mathbf{F}_{\text{cam}}^{\text{global: } t\to t'} - \mathbf{F}_{\text{est}}^{\text{t}\to\text{t}'})\|_1, \tag{10}$$

where $\cdot$ indicates element-wise multiplication. Note that the confident dynamic mask is initialized using the foundation models as described in Sec. 3.3. During the optimization, we use the global static pointmaps and camera parameters to compute $\mathbf{F}_{\text{cam}}^{\text{global}}$ and update the confident dynamic mask.

A.3 Ablation Study on Different Components for Dynamic Bundle Adjustment

Our dynamic BA pipeline introduces three key components absent in prior work like Uni4D [55], which systematically improve the decomposition of static/dynamic elements and global consistency:

- **(a) Epi-Mask-Based Dynamics Filtering:** We introduce a geometric filtering step using an epipolar-based mask ("Epi-mask") to achieve a cleaner separation between static background and dynamic foreground pixels before bundle adjustment. This leads to more stable camera pose estimation and background reconstruction.
- **(b) VLM-Based Semantic Dynamics Analysis:** We leverage a Vision-Language Model (VLM) for a high-level, semantic understanding of motion. This enables intelligent, motion-aware keyframe extraction and provides robust masks for dynamic objects, a significant improvement over purely geometric or flow-based segmentation.
- **(c) Optical Flow-Based Sliding Window Global Refinement:** To address error accumulation and temporal drift common in long videos, we implement a global refinement strategy over a sliding window. This enforces long-range temporal consistency, correcting errors that a frame-by-frame or local BA approach would miss.

Table 6: Components Ablation on Sintel.

Ablations	**(a)**	**(b)**	**(c)**	**ATE**↓	**RPE$_{trans}$**↓	**RPE$_{rot}$**↓	**Abs**↓	δ**1.25**↑
Baseline				0.114694	0.032125	0.347920	0.216433	0.725167
Ablation-1	✓			0.114065	0.032250	0.335198	0.215058	0.726943
Ablation-2		✓		0.11053	0.033122	0.334005	0.210339	0.722999
Ablation-3			✓	0.114694	0.032125	0.347920	0.214282	0.724084
Ablation-4	✓	✓		0.108459	0.028906	0.281979	0.205892	0.727616
Ablation-5	✓		✓	0.114065	0.032250	0.335198	0.214143	0.725534
Ablation-6		✓	✓	0.110530	0.033122	0.334005	0.207329	0.725784
DynamicGen (Ours)	✓	✓	✓	**0.108459**	**0.028906**	**0.281979**	**0.204574**	**0.728961**

A.4 Additional experiments on generated hierarchical captions.

We performed three distinct experiments to validate the high quality of our hierarchical semantic annotations:

- **(a) Object-Level Semantics via 4D-LangSplat [67]:** To validate the annotations produced by our DynamicGen framework, we performed a time-sensitive querying experiment using a 4D-LangSplat model. For this evaluation, we trained the model on the "americano" scene from the HyperNeRF dataset and benchmarked it against a re-implemented 4D-LangSplat* baseline. The results, presented in Tab. 7, demonstrate that our approach yields substantial gains in Accuracy and volumetric Intersection over Union (vIoU). This superior performance confirms that our precise object masks and labels are highly effective for demanding multi-modal applications.

Table 7: Quantitative comparisons of time-sensitive querying on the HyperNeRF [86] dataset.

Method	americano	
	Acc(%)	vIoU(%)
4D-LangSplat* [67]	53.84	27.55
DynamicGen	**64.42**	**51.65**

- **(b) Scene-Level Semantics via G-VEval [81]:** To rigorously assess our scene-level captions, we moved beyond single-score metrics and employed a more granular evaluation using the ACCR framework in G-VEval benchmark. This approach provides a comprehensive, multi-dimensional assessment of caption quality across four key axes: Accuracy, Completeness, Conciseness, and Relevance. On a random sample of 100 videos from SA-V data, our generated captions demonstrated high performance across all four criteria, as detailed in the Tab. 8. The strong

 https://doi.org/10.52202/085713-3621

performance across these metrics confirms that our captions are not only factually accurate and relevant to the video content, but also complete in their coverage of events and efficiently concise. This robust, multi-faceted quality makes them highly suitable and reliable for demanding downstream applications.

Table 8: Evaluation of generated captions using the ACCR framework from G-VEval.

Evaluation Criteria	Accuracy↑	Completeness↑	Conciseness↑	Relevance↑	Average↑
Scene-Level Captions	84.38	82.09	75.87	85.56	81.97

- **(c) Camera-Level Semantics via Human Study:** We conducted a formal human study to quantitatively analyze the quality of the final camera motion captions. Following prior work [71], we asked human evaluators to rate our captions on three criteria: (1) Clearness (clarity of information), (2) Conciseness (brevity without losing clarity), and (3) Grammar & Fluency. On a sub-sample of 88 videos from our dataset (i.e., filtered DAVIS), our captions performed excellently. The results, presented in Tab. 9 showed that over 60.22% of the captions were rated as both clear and fluent, while also receiving high scores for conciseness. This confirms the effectiveness of our generation and quality control process.

Table 9: Human evaluation results for the generated camera captions. Scores indicate the percentage of captions that met each quality criterion.

Human Evaluation	Rated as Clear	Rated as Fluent	Rated as Concise
Camera Captions	85.22%	89.77%	67.04%

A.5 More qualitative results of dynamic bundle adjustment

We present additional qualitative reconstruction results in Fig. 8, demonstrating the generalizability and performance of our pipeline on real-world data.

A.6 Inference Speed and Computational Cost for DynamicGen

For a reproducible analysis of computational performance, we processed the entire Sintel training set (23 videos) on NVIDIA H20 GPUs. A detailed breakdown of the average processing time and peak VRAM consumption for each component of our pipeline is provided in Table 10.

Table 10: Computational Cost Analysis.

Module	Hardware Used	Avg. Time / Sintel Video (mins)	Peak VRAM (GB)	Notes
1. Motion-aware Keyframe Extraction	1x H20 GPU	∼0.1	∼10	Selects representative frames
2. VLM-Based Semantic Analysis (Qwen-VL)	2x H20 GPU	∼1.6	∼60	Identifies dynamic elements
3. Moving Object Segmentation (SA2VA)	1x H20 GPU	∼0.8	∼30	Per-object video segmentation
4. Dynamic Bundle Adjustment	1x CPU Core + 1x H20 GPU	∼12.2	∼30	Main time bottleneck
5. Moving Object Captioning	2x H20 GPU	∼2.0	∼24	Object-level descriptions
6. Dynamic Scene Captioning	2x H20 GPU	∼3.0	∼40	Scene-level descriptions
7. Camera Motion Captioning	2x H20 GPU	∼2.0	∼40	Camera-level descriptions
8. Caption Rephrasing	1x H20 GPU	∼2.0	∼24	LLM-based refinement for consistency and conciseness
Total (per video)	**H20 GPU**	**∼23.7**	**∼60**	**Peak VRAM, not sum**

A.7 Limitations

Despite its considerable capabilities, DynamicVerse exhibits several inherent limitations. First, its reliance on in-the-wild internet videos introduces significant noise and quality variance. This can compromise the fidelity of metric-scale geometry and motion recovery, particularly in complex, cluttered, or occluded scenes that fall outside the typical distribution of the foundation models' training

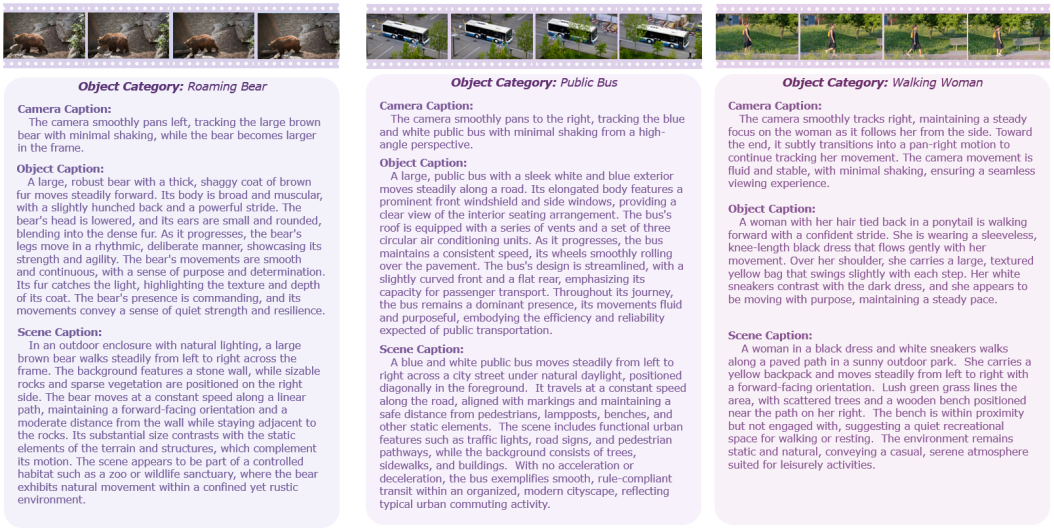


Figure 7: Examples captions on DAVIS dataset.

data. Second, the substantial computational overhead required to process long video sequences with large-scale models presents a practical barrier to real-time performance and scalable deployment. Finally, while extensive, the dataset cannot exhaustively capture the long tail of real-world phenomena. Consequently, the model's generalization to truly novel environments is fundamentally tethered to the intrinsic biases and capabilities of its underlying foundation models.

These limitations raise AI-safety concerns: (i) privacy and security risks, since metric-scale reconstructions from web videos can expose sensitive interiors or critical infrastructure and facilitate covert mapping or surveillance; and (ii) miscalibrated confidence under distribution shift, producing plausible but erroneous geometry and dynamics that misguide downstream robotic or AR planners. Biases and licensing gaps in foundation models and web data may further perpetuate representational harms and legal or IP issues. A practical mitigation is to prefilter ineligible videos using policy rules and automated detectors (e.g., content with PII, sensitive interiors or infrastructure, minors, or restricted licenses).

https://doi.org/10.52202/085713-3621

Figure 8: **Qualitative Results on in-the-wild data**. We show qualitatively some of our reconstruction results on in-the-wild data. For full reconstruction, please refer to our attached supplementary webpage.

Figure 9: **Qualitative Results of moving object Segmentation**. We show qualitatively some of our segmentation results on the Youtube-VIS dataset compared with other baselines.

https://doi.org/10.52202/085713-3621