TOWARDS CONSISTENT VIDEO GEOMETRY ESTIMATION

Anonymous authors
Paper under double-blind review

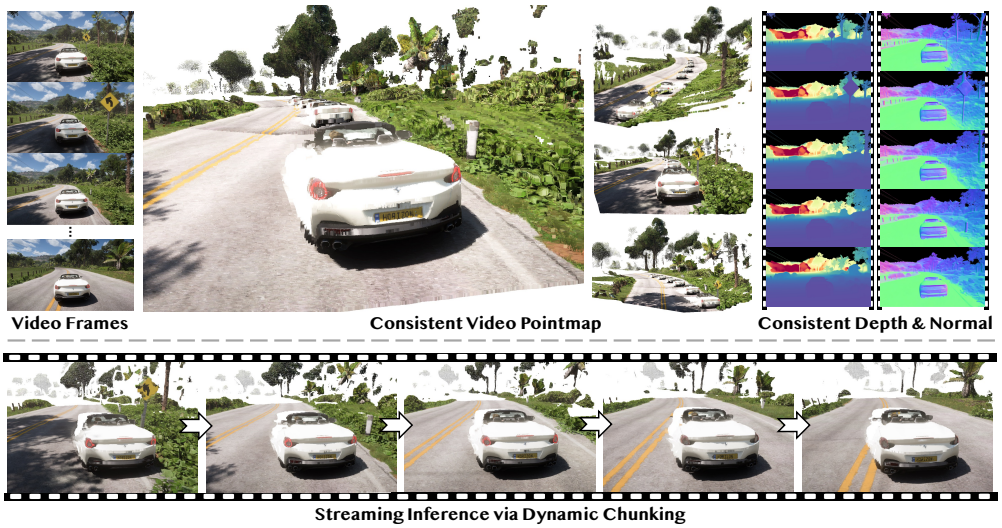

Figure 1: **ViGeo** is a unified feed-forward foundation model for video geometry estimation. It predicts temporally consistent depth, surface normals, and dense point maps from raw video frames. With dynamic chunking attention, the same trained model seamlessly switches between full-sequence reconstruction and streaming inference without retraining.

ABSTRACT

This work presents **ViGeo**, a feed-forward foundation model for recovering spatially dense and temporally consistent geometry from video sequences. Built upon a plain transformer architecture without task-specific architectural modifications, ViGeo supports streaming, full-sequence, and long-video inference within a unified model. The key design is **dynamic chunking attention**, which exposes the model to both bidirectional and causal temporal contexts during training and allows it to adapt its attention pattern at test time without retraining. To improve supervision quality, we further introduce a **completion-based data refinement framework**. This framework trains a video depth completion teacher that conditions on sparse and noisy annotations and exploits video/multi-view context to produce dense, temporally coherent, and geometrically reliable training targets. Beyond depth and point maps, ViGeo also predicts surface normals within the same framework. Trained solely on public datasets, ViGeo achieves state-of-the-art performance across online, offline, and long-video depth estimation, surface normal estimation, and video point map estimation.

1 INTRODUCTION

Video geometry estimation is a fundamental problem in computer vision, supporting applications such as robotics Liu et al. (2025), augmented reality Valentin et al. (2018), autonomous navigation Angelopoulos et al. (2019), and video editing Ceylan et al. (2023). These applications require geometry that is both spatially accurate and temporally consistent over long video sequences. Despite

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

recent progress, achieving high-fidelity reconstruction, long-term consistency, and scalable inference within a unified video model remains challenging.

A central limitation of existing video geometry models is their fixed temporal access pattern: offline methods Wang et al. (2025a; 2026b); Gao et al. (2025); Keetha et al. (2026) rely on future frames for full-sequence reasoning, while online methods Zhuo et al. (2025); LAN et al. (2026); Wang et al. (2025c); Wang & Agapito (2024) operate with restricted causal context. As a result, current models cannot adapt their attention behavior to the available video context at inference time. Large-scale training supervision poses another bottleneck: real-captured video geometry datasets are commonly built from LiDAR measurements Sun et al. (2020); Yeshwanth et al. (2023); Baruch et al. (2021) or SfM reconstructions Ling et al. (2024); Schonberger & Frahm (2016), whose sparse, noisy, or scale-ambiguous annotations limit spatial sharpness and temporal consistency.

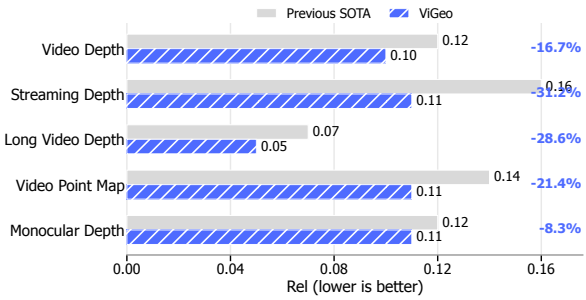

Figure 2: Benchmark comparison with previous state-of-the-art methods.

In this work, we present **ViGeo**, a feed-forward foundation model for dense and temporally consistent geometry estimation from video sequences. Instead of using separate architectures or training protocols for different inference regimes, ViGeo adopts a plain transformer backbone with **dynamic chunking attention**. This design exposes the model to both bidirectional and causal temporal contexts during training, allowing it to adapt its attention pattern at inference time without retraining. By changing the chunk partition, ViGeo can switch between full-sequence and streaming inference, while enabling scalable long-video processing by combining chunk-based inference with key-value (KV) caching Yuan et al. (2026).

To improve supervision from real-captured data, we further introduce a **completion-based data refinement framework** for scalable video geometry learning. Rather than treating raw annotations as reliable ground truth, we view them as imperfect geometric observations that should be completed and rectified. Prior refinement pipelines often rely on monocular depth prediction, followed by either affine alignment to sparse observations Yang et al. (2024b;c); Lin et al. (2026) or reconstruction-based post-processing Wang et al. (2025e). In contrast, our framework trains a video depth completion teacher that conditions on sparse and noisy annotations while leveraging temporal and multi-view context to produce dense, temporally coherent, and geometrically reliable training targets. This refinement process can be applied across diverse real-captured datasets, providing a practical data engine for large-scale video geometry supervision.

ViGeo also supports surface normal estimation alongside depth and point map prediction within the same framework. To reflect the practical requirements of video geometry estimation, we evaluate ViGeo across streaming, offline, and long-video depth estimation, as well as surface normal and point map estimation. Trained solely on publicly available datasets, ViGeo achieves state-of-the-art results on most metrics and remains competitive on the rest.

Our contributions are summarized as follows:

1. We present **ViGeo**, a feed-forward foundation model for dense and temporally consistent video geometry estimation. Built upon a plain transformer backbone, ViGeo supports depth, surface normal, and point map estimation within a unified framework.

2. We introduce **dynamic chunking attention**, which exposes the model to both bidirectional and causal temporal contexts during training. This design enables a single trained model to switch between full-sequence and streaming inference without retraining. Combined with KV caching, its chunk-based mode scales to long videos.

3. We propose a **completion-based data refinement framework** that trains a video depth completion teacher to refine sparse and noisy LiDAR/SfM annotations into dense, temporally coherent, and geometrically reliable training targets.

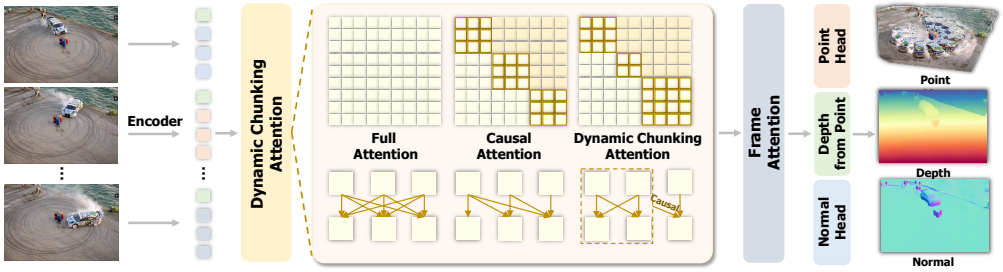

Figure 3: Architecture overview of **ViGeo**. Built upon a plain Transformer with dynamic chunking attention, ViGeo supports full-sequence, streaming, and long-video inference within a unified model and predicts temporally consistent depth, surface normals, and point maps.

4. We conduct extensive evaluations across multiple datasets and benchmarks, covering streaming, offline, and long-video depth estimation, as well as surface normal and point map estimation. ViGeo achieves state-of-the-art performance and demonstrates strong generalization across diverse video geometry settings.

2 METHOD

This section presents the methodology of **ViGeo**, a unified feed-forward framework for consistent monocular video geometry estimation. We first describe the overall network architecture in Sec. 2.1. Sec. 2.2 then introduces dynamic chunking attention, which enables a single trained model to adapt to streaming, full-sequence, and long-video inference without retraining. Next, Sec. 2.3 describes our completion-based data refinement framework, which trains a video depth completion teacher to construct dense and temporally coherent supervision from sparse and noisy real-world annotations. Training objectives and implementation details are provided in Appendix B.1.

2.1 OVERALL ARCHITECTURE

Given a video clip of N RGB frames $\{\mathbf{I}_i\}_{i=1}^N$, ViGeo predicts dense geometric quantities for each frame in a fully feed-forward manner. As illustrated in Fig. 3, ViGeo is built upon a plain ViT-style Transformer. The early layers operate within individual frames to extract dense visual tokens, while the later layers alternate between intra-frame attention and dynamic chunking attention to jointly model spatial details and temporal dependencies. The resulting spatiotemporal features are then decoded into point maps, depth maps, and surface normals. Formally, ViGeo maps the input sequence to a set of per-frame geometric predictions:

$$\{\mathbf{P}_i, \mathbf{D}_i, \mathbf{N}_i\}_{i=1}^N = f\big(\{\mathbf{I}_i\}_{i=1}^N\big), \tag{1}$$

where $\mathbf{P}_i \in \mathbb{R}^{3\times H\times W}$ denotes the point map, $\mathbf{D}_i \in \mathbb{R}^{H\times W}$ denotes the depth map, and $\mathbf{N}_i \in \mathbb{R}^{3\times H\times W}$ denotes the surface normal map of frame i.

Our architecture follows the recent trend of large feed-forward geometric models, but is designed for a more flexible video inference setting. Instead of committing to either full-sequence attention Wang et al. (2025a; 2026b); Lin et al. (2026) or causal attention Zhuo et al. (2025); LAN et al. (2026), ViGeo employs dynamic chunking attention to bridge these temporal access patterns within a single model. Together with intra-frame attention, this design keeps the backbone simple and generic, while allowing the same trained model to operate under offline, streaming, and long-video inference without architectural modification or retraining.

2.2 DYNAMIC CHUNKING ATTENTION DESIGN

Existing video geometry models usually adopt either full-sequence bidirectional attention for offline reconstruction Wang et al. (2025a; 2026b); Lin et al. (2026); Gao et al. (2025) or causal attention for streaming inference Zhuo et al. (2025); LAN et al. (2026). Instead of fixing the temporal access pattern, we introduce **dynamic chunking attention**, which allows a single model to adapt its attention

Table 1: Inference modes induced by dynamic chunking attention. By changing only the temporal chunk partition, the same attention formulation covers full-sequence, streaming, and chunk-based inference while sharing the same model parameters.

Mode	Chunk	Intra-Chunk	Inter-Chunk
Full-sequence (Offline)	$[\mathcal{N}]$	Full attention	–
Chunk-based	$[\mathcal{N}_1, \ldots, \mathcal{N}_L]$	Full attention	Causal attention
Streaming (Online)	$[1, \ldots, 1]$	–	Causal attention

behavior through the chunk partition. Given a sequence of frames, tokens attend bidirectionally within the same chunk and causally across different chunks. In other words, attention is full within each chunk and causal across chunks.

Formally, we partition the input sequence into a set of contiguous temporal chunks:

$$\mathcal{N} = \{\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_L\}, \qquad \sum_{l=1}^{L} |\mathcal{N}_l| = N, \tag{2}$$

where $|\mathcal{N}_l|$ denotes the number of consecutive frames in the l-th chunk, and N is the total sequence length. Let $\mathrm{ch}(i)$ denote the chunk index of frame i. We define a frame-level attention mask $\mathcal{M}^{\text{attn}}$, where the entry between query frame i and key frame j is given by:

$$\mathcal{M}^{\text{attn}}_{i,j} = \begin{cases} 1, & \mathrm{ch}(j) \le \mathrm{ch}(i), \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This mask is applied to all visual tokens according to their frame indices. When two frames belong to the same chunk, they can attend to each other bidirectionally. When they belong to different chunks, a frame can only attend to frames from previous chunks.

As summarized in Table 1, different chunk partitions instantiate different inference modes under the same formulation. When $L = 1$, all frames belong to a single chunk, and Eq. 3 reduces to full-sequence bidirectional attention for offline inference. When $|\mathcal{N}_l| = 1$ for all l, each chunk contains one frame and the mask becomes strictly causal, enabling streaming inference. Intermediate chunk sizes induce chunk-based inference, preserving bidirectional context within local temporal groups while maintaining causal access across chunks. During training, we expose the model to multiple chunk configurations, including both bidirectional and causal temporal contexts. At inference time, the same trained model can switch among full-sequence, streaming, and chunk-based inference by specifying the chunk partition, without modifying the architecture or retraining.

Dynamic chunking attention also supports scalable long-video processing. For long sequences, chunk-based inference is compatible with KV caching Yuan et al. (2026), allowing past states to be reused across chunks and helping control memory growth. This formulation also fits practical streaming scenarios, where inputs may arrive in short multi-frame packets.

2.3 Completion-Based Data Refinement

In practice, real-captured depth annotations Ling et al. (2024); Xia et al. (2024); Yao et al. (2020); Baruch et al. (2021) often contain missing regions, outliers, and scale ambiguity. Directly using such measurements as supervision can degrade spatial fidelity and temporal consistency. Prior refinement pipelines Wang et al. (2025e); Lin et al. (2026); Yang et al. (2024b;c) often rely on monocular depth predictions, followed by alignment to sparse observations or reconstruction-based post-processing. In contrast, we formulate real-data supervision refinement as a video depth completion problem. Our pipeline treats raw annotations as imperfect geometric observations and trains a video depth completion teacher, which is then used to produce dense, temporally coherent, and geometrically reliable pseudo-labels. As shown in Fig. 4, the pipeline consists of two stages: per-frame outlier filtering and multi-frame video depth completion.

Outlier Filtering. We first filter unreliable raw measurements before depth completion. Given a raw depth map $\mathbf{D}^{\text{raw}}$, we use the local spherical alignment criterion from MoGe-2 Wang et al. (2025e) to

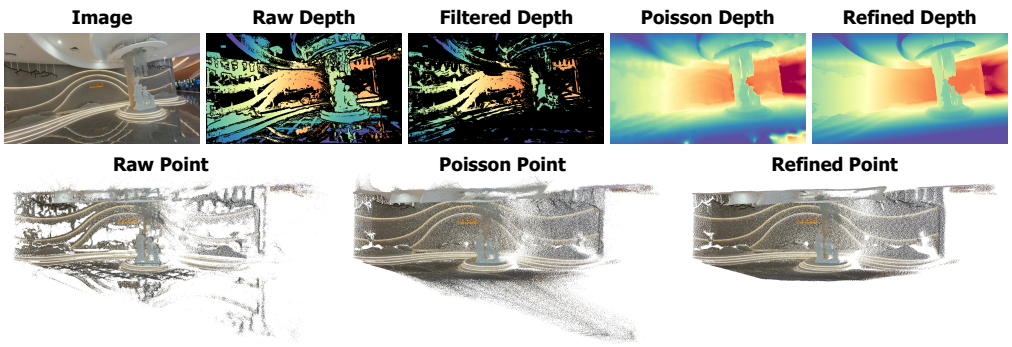

Figure 4: Visualization of our completion-based data refinement pipeline. Given an RGB video sequence and raw depth maps with missing regions and noisy measurements, we first apply per-frame outlier filtering to obtain reliable sparse observations, and then construct dense but coarse depth priors via Poisson reconstruction. A video depth completion teacher refines these priors by leveraging temporal and multi-view context, producing sharp and coherent dense depth labels. The bottom row compares the corresponding point clouds, where our refined labels reduce missing regions, flying points, and geometric artifacts.

identify inconsistent observations, yielding a valid mask $\mathcal{M}^{\text{valid}}$. The filtered sparse depth is obtained as $\tilde{\mathbf{D}} = \mathbf{D}^{\text{raw}} \odot \mathcal{M}^{\text{valid}}$, where $\odot$ denotes the Hadamard product.

Video Depth Completion Teacher. Given the filtered sparse depth sequence, we use the trained video depth completion teacher to generate dense pseudo-labels. To provide a dense geometric condition for the teacher, we first convert the filtered sparse depth $\tilde{\mathbf{D}}$ into a coarse dense prior $\mathbf{D}^{\text{prior}}$ using Poisson reconstruction Yu et al. (2026). Following LDCM Yu et al. (2026), the prior is obtained by aligning its log-gradient field with that of an initial monocular relative depth prediction $\mathbf{D}^{\text{mono}}$, while preserving the reliable sparse measurements in $\tilde{\mathbf{D}}$:

$$\min_{\mathbf{D}^{\text{prior}}} \sum_{p} \left\| \nabla \log \mathbf{D}_p^{\text{prior}} - \nabla \log(\mathbf{D}^{\text{mono}} + \gamma)_p \right\|^2 + \lambda \sum_{p \in \mathcal{M}^{\text{valid}}} \left(\mathbf{D}_p^{\text{prior}} - \tilde{\mathbf{D}}_p \right)^2, \tag{4}$$

where γ is a shift factor derived from affine alignment Yu et al. (2026). Unlike prior pipelines that directly use the monocular prediction or its aligned reconstruction as the final supervision, this prior only serves as a dense geometric condition for the video completion teacher.

Given the full sequence of RGB images $\{\mathbf{I}_i\}_{i=1}^N$ and the corresponding dense priors $\{\mathbf{D}_i^{\text{prior}}\}_{i=1}^N$, we apply median-based log normalization to handle scale-ambiguous data Ling et al. (2024); Yao et al. (2020), such as SfM reconstructions. Specifically, the dense priors are normalized as $\log(\mathbf{D}_i^{\text{prior}}/m)$, where m is the median depth value computed over valid sparse measurements across the temporal sequence. Following LingBot-Depth Tan et al. (2026), RGB images and normalized depth priors are separately embedded as patch tokens with spatial and modality-specific positional encodings.

The teacher adopts a ViT-style architecture similar to ViGeo, extending this RGB-prior completion formulation from single images to video sequences. Its deeper layers aggregate intra-frame and cross-frame context to complete the coarse priors into temporally coherent dense depth predictions. The predicted depth is restored to the original scale by multiplying it with the sequence median m, yielding the final dense pseudo-labels:

$$\{\hat{\mathbf{D}}_i^{\text{pseudo}}\}_{i=1}^N = \mathcal{F}^{\text{teacher}} \left(\{\mathbf{I}_i\}_{i=1}^N, \left\{ \log(\mathbf{D}_i^{\text{prior}}/m) \right\}_{i=1}^N \right) \cdot m. \tag{5}$$

These dense pseudo-labels $\{\hat{\mathbf{D}}_i^{\text{pseudo}}\}_{i=1}^N$ replace the raw measurements as supervision for training ViGeo on real-captured data.

Fig. 4 illustrates the effect of each refinement stage. Raw measurements contain missing regions and outliers. Poisson reconstruction densifies the depth but may introduce flying points and geometric artifacts. The video depth completion teacher further refines these priors, producing denser and more coherent point clouds that better align with image structures. Additional qualitative examples are provided in Appendix E.

Table 2: **Quantitative results for scale-invariant video depth estimation** on the Sintel Bozic et al. (2021), Bonn Palazzolo et al. (2019), and KITTI Geiger et al. (2013) datasets. Baselines are grouped into offline and online settings and evaluated via absolute relative error (Rel) and threshold accuracy (δ_1). The best and second-best results in each category are highlighted in **bold** and underlined, respectively.

Method	Sintel		Bonn		KITTI	
	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑
Offline						
VDA Chen et al. (2025)	0.855	0.340	0.301	0.465	0.272	0.535
DepthCrafter Hu et al. (2024b)	0.469	0.455	0.074	0.968	0.282	0.435
GeometryCrafter Xu et al. (2025b)	0.305	0.675	0.057	0.969	0.057	0.981
VGGT Wang et al. (2025a)	0.373	0.634	0.057	0.966	0.062	0.969
Pi3 Wang et al. (2026b)	0.281	0.663	0.049	0.975	0.038	**0.985**
DA3 Lin et al. (2026)	0.284	0.666	0.050	0.972	0.058	0.976
ViGeo (offline)	**0.229**	**0.734**	**0.046**	**0.976**	**0.037**	0.982
Online						
StreamVGGT Zhuo et al. (2025)	0.407	0.614	0.059	0.971	0.173	0.721
Stream3R LAN et al. (2026)	0.319	0.675	0.068	0.956	0.088	0.938
FlashDepth Chou et al. (2025)	0.361	0.562	0.066	0.974	0.064	0.971
ViGeo (online)	**0.242**	**0.697**	**0.051**	**0.974**	**0.048**	**0.976**

3 Experiments

Detailed experimental settings, including datasets, evaluation protocols, and implementation details, are provided in Appendices A and B.

Table 3: **Quantitative results for scale-invariant long video depth estimation** on the Bonn Palazzolo et al. (2019), KITTI Geiger et al. (2013), and HAMMER Jung et al. (2023) datasets. We report the absolute relative error (Rel) and threshold accuracy (δ_1). The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Bonn (400)		KITTI (300)		HAMMER (300)	
	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑
VGGT Wang et al. (2025a)	OOM	OOM	OOM	OOM	OOM	OOM
StreamVGGT Zhuo et al. (2025)	OOM	OOM	OOM	OOM	OOM	OOM
Stream3R LAN et al. (2026)	OOM	OOM	OOM	OOM	OOM	OOM
VDA Chen et al. (2025)	0.291	0.531	0.274	0.465	0.278	0.482
DepthCrafter Hu et al. (2024b)	0.132	0.891	0.288	0.451	0.490	0.302
GeometryCrafter Xu et al. (2025b)	0.101	0.934	0.070	0.970	0.054	0.983
FlashDepth Chou et al. (2025)	0.084	0.943	0.100	0.919	0.064	0.971
InfiniteVGGT Yuan et al. (2026)	0.069	0.965	0.206	0.659	0.093	0.921
ViGeo	**0.059**	**0.967**	**0.059**	**0.972**	**0.023**	**0.997**

3.1 Video Depth Estimation

We evaluate ViGeo for video depth estimation on the Sintel Bozic et al. (2021), Bonn Palazzolo et al. (2019), and KITTI Geiger et al. (2013) datasets, with quantitative results summarized in Table 2. To comprehensively assess our model, we compare it against a diverse set of baselines that encompass both offline and online paradigms. These include prominent video depth estimators (VideoDepthAnything Chen et al. (2025), DepthCrafter Hu et al. (2024b), FlashDepth Chou et al. (2025)), as well as state-of-the-art 3D reconstruction and depth foundation models (VGGT Wang

Table 4: **Quantitative results for scale-invariant video point map estimation** on the Sintel Bozic et al. (2021), Bonn Palazzolo et al. (2019), and KITTI Geiger et al. (2013) datasets. Baselines are grouped into offline and online settings and evaluated via point-wise relative error (Rel^p) and threshold accuracy ($\delta^p_{0.25}$) computed over the 3D point coordinates. The best and second-best results in each category are highlighted in **bold** and underlined, respectively.

Method	Sintel		Bonn		KITTI	
	$\text{Rel}^p \downarrow$	$\delta^p_{0.25} \uparrow$	$\text{Rel}^p \downarrow$	$\delta^p_{0.25} \uparrow$	$\text{Rel}^p \downarrow$	$\delta^p_{0.25} \uparrow$
Offline						
GeometryCrafter Xu et al. (2025b)	0.280	0.642	0.070	0.975	0.084	0.984
VGGT Wang et al. (2025a)	0.378	0.307	0.129	0.972	0.358	0.230
Pi3 Wang et al. (2026b)	0.257	0.661	0.097	0.821	0.073	**0.992**
DA3 Lin et al. (2026)	0.285	0.589	0.059	0.975	0.148	0.935
ViGeo (offline)	**0.223**	**0.763**	**0.053**	**0.978**	**0.054**	0.986
Online						
StreamVGGT Zhuo et al. (2025)	0.402	0.304	0.129	0.974	0.403	0.142
Stream3R LAN et al. (2026)	0.386	0.266	0.146	0.961	0.365	0.211
ViGeo (online)	**0.237**	**0.711**	**0.060**	**0.977**	**0.067**	**0.980**

et al. (2025a), Pi3 Wang et al. (2026b), Depth Anything 3 Lin et al. (2026), CUT3R Wang et al. (2025c), StreamVGGT Zhuo et al. (2025), Stream3R LAN et al. (2026)). Following standard practice, all predicted depth sequences are aligned to the ground truth using a single scale factor per sequence. We report the absolute relative error (Abs Rel) and threshold accuracy (δ_1). The results demonstrate the robust effectiveness of our method across various scenarios, maintaining consistently high performance in both settings. Notably, even when restricted to the online streaming mode, ViGeo surpasses several existing offline methods.

3.2 Long Video Depth Estimation

To evaluate the long video depth estimation capability of ViGeo, we benchmark our model on Bonn Palazzolo et al. (2019), KITTI Geiger et al. (2013), and HAMMER Jung et al. (2023) with extended sequences of 300–400 frames. As shown in Table 3, several strong baselines, including VGGT Wang et al. (2025a) and its streaming variants Zhuo et al. (2025); LAN et al. (2026), encounter out-of-memory (OOM) errors and fail to process these long sequences. In contrast, ViGeo handles long videos through dynamic chunking attention, which adapts the temporal access pattern at inference time and remains compatible with key-value (KV) caching Yuan et al. (2026). Quantitative results show that ViGeo achieves state-of-the-art performance across all three datasets, outperforming both offline and online baselines. Its consistently low relative error and high threshold accuracy (δ_1) indicate strong temporal stability and robustness on extended video sequences.

3.3 Video Point Map Estimation

To assess the ability of ViGeo to recover dense 3D geometry, we evaluate video point map estimation on the same datasets used for video depth estimation. Since point maps represent per-pixel 3D geometry, we report the point-wise relative error (Rel^p) and threshold accuracy ($\delta^p_{0.25}$) over 3D point coordinates. As shown in Table 4, ViGeo achieves strong geometric accuracy and consistently outperforms existing baselines across all evaluated datasets. These results demonstrate that ViGeo recovers accurate dense 3D geometry in video sequences.

3.4 Ablation Study

Dynamic Chunking Attention. To validate the effectiveness of dynamic chunking attention, we train three ViGeo variants with different temporal attention schemes: full-sequence attention, causal attention, and the proposed dynamic chunking attention. All variants are evaluated under both offline full-sequence inference and online streaming inference. As shown in Table 5, full-attention

Table 5: **Ablation study on model architecture.** We evaluate the versatility of ViGeo by comparing variants trained with different attention strategies. All models are evaluated under both offline (full-sequence) and online (streaming) inference modes on regular video benchmarks.

Training Scheme	Sintel		Bonn		KITTI	
	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑
Offline						
Full Attention	**0.289**	0.626	**0.051**	0.971	**0.047**	**0.978**
Causal Attention	0.328	0.589	0.058	0.966	0.059	0.971
Dynamic Chunking (Ours)	0.291	**0.627**	**0.051**	**0.973**	0.048	**0.978**
Online						
Full Attention	0.363	0.551	0.072	0.955	0.148	0.699
Causal Attention	0.330	0.560	0.066	**0.963**	0.057	**0.972**
Dynamic Chunking (Ours)	**0.322**	**0.577**	**0.065**	**0.963**	**0.055**	**0.972**

Table 6: **Ablation study on data refinement framework.** We compare the performance of models trained using raw sensor measurements versus our refined pseudo-labels. All variants are evaluated on the Sintel, Bonn, and KITTI benchmarks.

Training Supervision	Sintel		Bonn		KITTI	
	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑
Raw Measurements	0.291	0.627	0.051	0.973	0.048	**0.978**
w/ Refined Labels (Ours)	**0.283**	**0.633**	**0.050**	**0.975**	**0.047**	**0.978**

training performs well in the offline setting but degrades noticeably when evaluated online, due to the mismatch between bidirectional training and causal inference. This degradation is particularly pronounced on KITTI, where fast camera motion and large scene depth ranges make the model more sensitive to changes in temporal context. In contrast, causal-attention training is better aligned with online inference, but it cannot fully exploit bidirectional context under offline evaluation. By exposing the model to both bidirectional and causal temporal contexts during training, dynamic chunking retains the strong offline performance of full-attention training while matching or outperforming causal-attention training online. It achieves the best or tied-best δ_1 on all three datasets offline and the best or tied-best results for every metric online, without retraining.

Completion-Based Data Refinement. To evaluate the impact of our data refinement framework, we compare ViGeo trained with raw measurements against the variant supervised by refined labels. As shown in Table 6, refined supervision improves most metrics under the same architecture and training protocol, indicating that higher-quality supervision benefits video geometry learning. This gain comes from the video depth completion teacher, which converts sparse and noisy observations into dense and temporally coherent training targets.

We further provide qualitative comparisons in Fig. 5. Raw sensor measurements often contain missing regions and outliers. Poisson reconstruction Yu et al. (2026); Wang et al. (2025e) densifies the observations but may introduce flying points and geometric distortions. In contrast, our full refinement pipeline produces cleaner and more complete point clouds with better structural coherence.

3.5 Qualitative Comparison

Qualitative comparisons of point cloud reconstruction are shown in Fig. 6. The results cover diverse scenarios, including outdoor scenes, indoor environments, and object-centric inputs. Compared with existing feed-forward approaches, ViGeo recovers cleaner and more complete 3D structures. Pi3 Wang et al. (2026b) often introduces structural noise and checkerboard artifacts, while VGGT Wang et al. (2025a) tends to produce incomplete or fragmented geometry. In contrast, ViGeo better preserves object shapes, scene layouts, and spatial consistency, leading to more coherent 3D reconstructions.

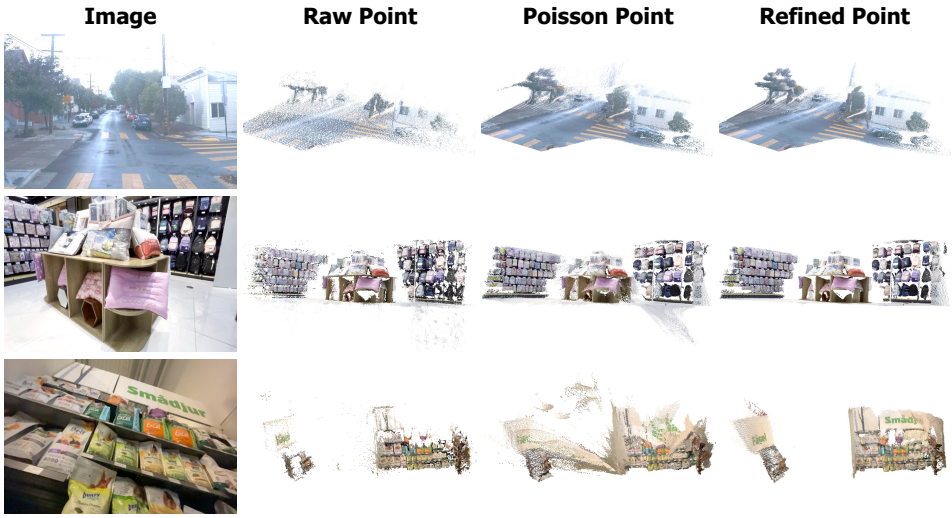

Figure 5: Qualitative ablation of our data refinement pipeline. Our full pipeline effectively resolves the severe missing regions in raw points and the flying points introduced by Poisson reconstruction, yielding clean and structurally coherent 3D geometries.

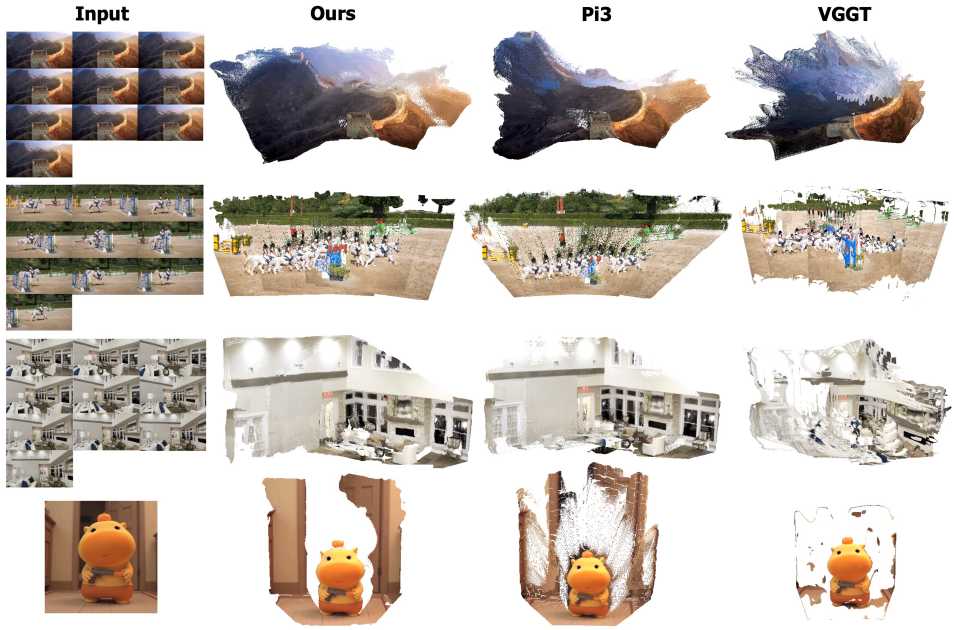

Figure 6: Qualitative results on 3D reconstruction. Our method yields more accurate and robust geometric structures across diverse scenarios when compared to existing feed-forward approaches.

4 Conclusion

In this paper, we present **ViGeo**, a feed-forward geometry foundation model for temporally consistent depth and surface normal estimation. By introducing **dynamic chunking attention**, ViGeo unifies streaming and full-sequence inference within a single transformer model. To improve supervision, we further develop a **robust data refinement framework** that converts sparse and noisy annotations into dense, coherent, and geometrically reliable targets. Experiments across multiple benchmarks show that ViGeo achieves state-of-the-art performance while maintaining strong spatial sharpness and temporal consistency. These results suggest that ViGeo provides a practical foundation for scalable video geometry estimation.

AI USE STATEMENT

In this work, generative AI tools were used to improve the English writing, assist with LaTeX and table formatting, review the manuscript structure and formatting, and check compliance with the conference submission guidelines. Generative AI tools were not used to formulate the research hypotheses, design the methodology or experiments, generate or process experimental data, implement the proposed method, or interpret the experimental results. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper, including all text, claims, results, and artifacts.

ETHICS STATEMENT

ViGeo reconstructs dense scene geometry from videos and may raise privacy concerns when applied to real-world recordings. It could also be misused for unauthorized mapping or surveillance. In addition, although the model is trained on publicly available datasets, it may inherit dataset biases and perform less reliably in underrepresented environments. Users should comply with applicable data licenses and privacy regulations, obtain appropriate authorization when processing real-world recordings, and carefully validate the model before deployment in safety-critical settings.

REFERENCES

Anastasios Nikolas Angelopoulos, Hossein Ameri, Debbie Mitra, and Mark Humayun. Enhanced depth navigation through augmented reality depth mapping in patients with low vision. *Scientific reports*, 9(1):11230, 2019.

Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9535–9545, 2024.

Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13137–13146, 2021.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4009–4018, 2021.

Yanrui Bin, Wenbo Hu, Haoyuan Wang, Xinya Chen, and Bing Wang. Normalcrafter: Learning temporally consistent normals from video diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2025.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8726–8737, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23206–23217, 2023.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22831–22840, 2025.

Gene Chou, Wenqi Xian, Guandao Yang, Mohamed Abdelfattah, Bharath Hariharan, Noah Snavely, Ning Yu, and Paul Debevec. Flashdepth: Real-time streaming video depth estimation at 2k resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9638–9648, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 27, 2014.

Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, and Chengfei Lyu. Dens3r: A foundation model for 3d geometry prediction. *arXiv preprint arXiv:2507.16290*, 2025.

Yi Feng, Zizhan Guo, Yu Ma, Hanli Wang, Rui Fan, et al. An instance-centric panoptic occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2603.27238*, 2026.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2002–2011, 2018.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Jingnan Gao, Zhe Wang, Xianze Fang, Xingyu Ren, Zhuo Chen, Shengqi Liu, Yuhao Cheng, Jiangjing Lyu, Xiaokang Yang, and Yichao Yan. More: 3d visual geometry reconstruction meets mixture-of-experts. *arXiv preprint arXiv:2510.27234*, 2025.

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.

Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9233–9243, 2023.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024.

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. *arXiv preprint arXiv:2404.15506*, 2024a.

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*, 2024b.

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

HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, et al. On the importance of accurate geometry data for dense 3d vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 780–791, 2023.

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 International Conference on Computer Vision*, 2023.

B Ke, K Qu, T Wang, N Metzger, S Huang, B Li, A Obukhov, and K Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv preprint arXiv:2505.09358*, 2025.

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*, pp. 9492–9502, 2024.

Nikhil Varma Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. Mapanything: Universal feed-forward metric 3d reconstruction. In *3DV*, 2026.

Yushi LAN, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. STream3r: Scalable sequential 3d reconstruction with causal transformer. In *International Conference on Learning Representations*, 2026.

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

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3205–3215, 2023.

LightwheelAI and LightwheelOcc contributors. Lightwheelocc: A 3d occupancy synthetic dataset in autonomous driving. https://github.com/OpenDriveLab/LightwheelOcc, 2024.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22160–22169, 2024.

Zeyi Liu, Shuang Li, Eric Cousineau, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Geometry-aware 4d video generation for robot manipulation. *arXiv preprint arXiv:2507.01099*, 2025.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22820–22830, 2025.

Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics*, 39(4):71–1, 2020.

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*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 7855–7862, 2019.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20133–20143, 2023.

Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. Tartanground: A large-scale dataset for ground robot perception and navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 20524–20531. IEEE, 2025.

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*, pp. 10106–10116, 2024.

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.

Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip HS Torr, Raquel Urtasun, and Jiaya Jia. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):969–984, 2020.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10912–10922, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3234–3243, 2016.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision*, pp. 746–760, 2012.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, et al. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5283–5293, 2025.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

Julien Valentin, Adarsh Kowdle, Jonathan T Barron, Neal Wadhwa, Max Dzitsiuk, Michael Schoenberg, Vivek Verma, Ambrus Csaszar, Eric Turner, Ivan Dryanovski, et al. Depth from motion for smartphone ar. *ACM Transactions on Graphics*, 37(6):1–19, 2018.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21686–21697, 2024a.

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*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, Xiangxiang Chu, and Yao Zhao. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338*, 2025b.

Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *arXiv preprint arXiv:1909.05452*, 2019.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10510–10522, 2025c.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5261–5271, 2025d.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, 2025e.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024b.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4909–4916, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026b.

Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9466–9476, 2023.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22378–22389, 2024.

Shaocong Xu, Songlin Wei, Qizhe Wei, Zheng Geng, Hong Li, Licheng Shen, Qianpu Sun, Shu Han, Bin Ma, Bohan Li, Chongjie Ye, Yuhang Zheng, Nan Wang, Saining Zhang, and Hao Zhao. Diffusion knows transparency: Repurposing video diffusion for transparent object depth and normal estimation. *arXiv preprint arXiv:2512.23705*, 2025a.

Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6632–6644, 2025b.

Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024a.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21924–21935, 2025.

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*, 2024b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024c.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1790–1799, 2020.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5684–5693, 2019.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9043–9053, 2023.

Zhu Yu, Zehua Sheng, Zili Zhou, Lun Luo, Si-Yuan Cao, Hong Gu, Huaqi Zhang, and Hui-Liang Shen. Aggregating feature point cloud for depth completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8732–8743, 2023.

Zhu Yu, zhengyi zhao, Runmin Zhang, Lingteng Qiu, Si-Yuan Cao, Kejie Qiu, Yisheng He, Siyu Zhu, Zilong Dong, and Hui liang Shen. Large depth completion model from sparse observations. In *International Conference on Learning Representations*, 2026.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. Infinitevggt: Visual geometry grounded transformer for endless streams, 2026.

Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3916–3925, 2022.

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. In *International Conference on Learning Representations*, 2025.

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*, pp. 19855–19865, 2023.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.

APPENDIX

This supplementary material provides additional details and analyses for **ViGeo**. The contents are organized as follows, where **bold** text highlights the primary sections.

Appendix A: **Datasets and Metrics**

- Training datasets and filtering details
- Benchmark details and evaluation protocols
- Evaluation metrics for depth, point maps, and surface normals

Appendix B: **Model Training**

- Training objectives
- Implementation details

Appendix C: **Related Work**

- Dense monocular geometry estimation
- Dense video geometry estimation
- Large-scale data training

Appendix D: **Additional Experimental Results**

- Surface normal estimation
- Monocular depth estimation
- Inference strategy for long-video processing
- Inference efficiency
- Additional long-video, variable-chunk, and streaming evaluations
- DCA post-training and completion-based refinement ablations
- Temporal consistency evaluation

Appendix E: **Additional Visualizations**

- Additional point cloud visualizations
- Additional qualitative results of ViGeo
- Completion-based data refinement results

Appendix F: **Limitations**

A DATASETS AND METRICS

A.1 TRAINING DATASETS

We collect 23 open-source RGB-D datasets to train ViGeo, comprising 17 synthetic and 6 real-world datasets. An overview of the training datasets is provided in Table 7, spanning five distinct domains: indoor, outdoor, driving, object, and in-the-wild scenarios. The combined training set covers a diverse range of environments, from controlled synthetic spaces to complex real-world captures utilizing LiDAR and 3D reconstruction techniques. The number of scenes and RGB-D pairs in each dataset may slightly differ from the originally released versions, as we manually filtered the data to exclude invalid frames and ensure high-quality training samples. Specifically, surface normal supervision is restricted to only high-quality synthetic datasets.

Table 7: An overview of the training datasets.

Dataset	Domain	#Scenes	Type
Hypersim Roberts et al. (2021)	Indoor	741	Synthetic
LightWheelOcc LightwheelAI & contributors (2024)	Outdoor/Driving	1020	Synthetic
TartanAir Wang et al. (2020)	In-the-wild	412	Synthetic
TartanGround Patel et al. (2025)	In-the-wild	8100	Synthetic
GTA-SfM Wang & Shen (2019)	Outdoor/In-the-wild	213	Synthetic
PointOdyssey Zheng et al. (2023)	Indoor	33	Synthetic
BEDLAM Black et al. (2023)	Indoor	7309	Synthetic
Dynamic Replica Karaev et al. (2023)	Indoor	1046	Synthetic
MatrixCity Li et al. (2023)	Outdoor/Driving	2581	Synthetic
MVS-Synth Huang et al. (2018)	Outdoor/Driving	120	Synthetic
OmniWorld Zhou et al. (2025)	In-the-wild	11424	Synthetic
Synthia Ros et al. (2016)	Outdoor/Driving	284	Synthetic
OmniObject3D Wu et al. (2023)	Object	4794	Synthetic
Aria Synthetic Environments Pan et al. (2023)	Indoor	22946	Synthetic
Spring Mehl et al. (2023)	In-the-wild	35	Synthetic
TransPhy3D Xu et al. (2025a)	Transparent/Reflective	11148	Synthetic
CarlaOcc Feng et al. (2026)	Outdoor/Driving	591	Synthetic
WildRGB Xia et al. (2024)	Indoor/Object	2058	LiDAR
Waymo Sun et al. (2020)	Outdoor/Driving	3790	LiDAR
ARKitScenes Baruch et al. (2021)	Indoor	4307	LiDAR
ScanNet++ Yeshwanth et al. (2023)	Indoor	1811	LiDAR
DL3DV Ling et al. (2024)	In-the-wild	6379	COLMAP
BlendedMVS Yao et al. (2020)	In-the-wild	503	3D Reconstruction

A.2 Evaluation Datasets

ViGeo is evaluated across multiple benchmark datasets and inference settings. Unless otherwise specified, input images are resized to satisfy each model's resolution requirements, and predictions are resized back to the original resolution before metric computation. We prepare the evaluation datasets as follows:

- **Sintel** Bozic et al. (2021): We use all sequences from the training split for evaluation. Each sequence contains 21 to 50 frames with an original resolution of 1024×436. Sintel is used for monocular depth estimation, video depth estimation, video point map estimation, and surface normal estimation. For depth evaluation, the maximum depth is set to 70 meters.
- **Bonn** Palazzolo et al. (2019): Following Pi3 Wang et al. (2026b), we evaluate on five Bonn sequences with an original resolution of 640×480. Bonn is used for monocular depth estimation, video depth estimation, long video depth estimation, and video point map estimation. For standard evaluation, we sample 110 frames from each sequence. For long video depth estimation, we sample 400 consecutive frames starting from the first frame of each sequence.
- **KITTI** Geiger et al. (2013): Following Pi3 Wang et al. (2026b), we evaluate on 13 KITTI sequences with an original resolution of 1242×375. KITTI is used for monocular depth estimation, video depth estimation, long video depth estimation, and video point map estimation. For standard evaluation, we sample 110 frames from each sequence. For long video depth estimation, we sample 300 consecutive frames from each sequence.
- **HAMMER** Jung et al. (2023): We evaluate on 11 HAMMER sequences with an original resolution of 1088×832. HAMMER is used for long video depth estimation and surface normal estimation. For long video depth estimation, we sample 300 consecutive frames from each sequence. For surface normal estimation, we sample 110 frames with a stride of 2.
- **NYUv2** Silberman et al. (2012): We use NYUv2 for monocular surface normal estimation and evaluate on the official test split of 654 images with an original resolution of 640×480. Following the standard evaluation protocol, images are cropped to 565×427 before metric computation.

A.3 Evaluation Metrics

We report evaluation metrics for depth estimation, video point map estimation, and surface normal estimation. All metrics are computed over valid pixels. We denote the set of valid frame-pixel pairs as $\mathcal{V}$. For scale-ambiguous predictions, we first align the prediction with a scalar scale factor s following the protocol in the main paper.

Depth metrics. For depth estimation, we report the absolute relative error Rel and threshold accuracy δ_1. Let $\mathbf{D}_{i,p}$ and $\hat{\mathbf{D}}_{i,p}$ denote the predicted and ground-truth depths at pixel p of frame i, respectively. The metrics are defined as:

$$\mathrm{Rel} = \frac{1}{|\mathcal{V}|} \sum_{(i,p)\in\mathcal{V}} \frac{\left| s\mathbf{D}_{i,p} - \hat{\mathbf{D}}_{i,p} \right|}{\hat{\mathbf{D}}_{i,p}}, \tag{6}$$

$$\delta_1 = \frac{1}{|\mathcal{V}|} \sum_{(i,p)\in\mathcal{V}} \mathbf{1}\left[\max\left(\frac{s\mathbf{D}_{i,p}}{\hat{\mathbf{D}}_{i,p}}, \frac{\hat{\mathbf{D}}_{i,p}}{s\mathbf{D}_{i,p}} \right) < 1.25 \right]. \tag{7}$$

Point map metrics. For video point map estimation, we report the point-wise relative error Rel^p and threshold accuracy $\delta^p_{0.25}$. Let $\mathbf{P}_{i,p}$ and $\hat{\mathbf{P}}_{i,p}$ denote the predicted and ground-truth point maps at pixel p of frame i, respectively. The metrics are defined as:

$$\mathrm{Rel}^p = \frac{1}{|\mathcal{V}|} \sum_{(i,p)\in\mathcal{V}} \frac{\left\| s\mathbf{P}_{i,p} - \hat{\mathbf{P}}_{i,p} \right\|_2}{\left\| \hat{\mathbf{P}}_{i,p} \right\|_2}, \tag{8}$$

$$\delta^p_{0.25} = \frac{1}{|\mathcal{V}|} \sum_{(i,p)\in\mathcal{V}} \mathbf{1}\left[\frac{\left\| s\mathbf{P}_{i,p} - \hat{\mathbf{P}}_{i,p} \right\|_2}{\left\| \hat{\mathbf{P}}_{i,p} \right\|_2} < 0.25 \right]. \tag{9}$$

Surface normal metrics. For surface normal estimation, we compute the angular error between the predicted normal $\mathbf{N}_{i,p}$ and ground-truth normal $\hat{\mathbf{N}}_{i,p}$:

$$\theta_{i,p} = \arccos\left(\frac{\mathbf{N}_{i,p}^\top \hat{\mathbf{N}}_{i,p}}{\|\mathbf{N}_{i,p}\|_2 \|\hat{\mathbf{N}}_{i,p}\|_2} \right), \tag{10}$$

We report the mean and median angular errors, denoted as Mean and Med, as well as $\delta_{11.25^\circ}$, the percentage of pixels whose angular error is below 11.25°:

$$\delta_{11.25^\circ} = \frac{1}{|\mathcal{V}|} \sum_{(i,p)\in\mathcal{V}} \mathbf{1}\left[\theta_{i,p} < 11.25^\circ \right]. \tag{11}$$

B Model Training

B.1 Training Objectives

We train ViGeo end-to-end with a multi-task geometry loss:

$$\mathcal{L} = \mathcal{L}_{\text{points}} + \lambda_{\text{points_normal}} \mathcal{L}_{\text{points_normal}} + \lambda_{\text{normal}} \mathcal{L}_{\text{normal}}. \tag{12}$$

Since depth is directly obtained from the predicted point map, we supervise the 3D point map as the primary geometric representation. The point map loss $\mathcal{L}_{\text{points}}$ penalizes the L_1 distance between the predicted and ground-truth point maps:

$$\mathcal{L}_{\text{points}} = \sum_{i=1}^{N} \sum_{p\in\mathcal{M}^{\text{valid}}} \frac{1}{\hat{\mathbf{D}}_{i,p}} \left\| s\mathbf{P}_{i,p} - \hat{\mathbf{P}}_{i,p} \right\|_1, \tag{13}$$

where $\mathbf{P}_{i,p}$ and $\hat{\mathbf{P}}_{i,p}$ denote the predicted and ground-truth point maps at pixel p of frame i, respectively, and $\hat{\mathbf{D}}_{i,p}$ denotes the ground-truth depth. To handle scale ambiguity, the scale factor s is estimated by minimizing:

$$s = \arg\min_{s'} \sum_{i=1}^{N} \sum_{p \in \mathcal{M}^{\text{valid}}} \frac{1}{\hat{\mathbf{D}}_{i,p}} \left\| s' \mathbf{P}_{i,p} - \hat{\mathbf{P}}_{i,p} \right\|_1, \tag{14}$$

which is efficiently solved using the ROE solver Wang et al. (2025d).

In addition to point-wise supervision, we impose surface geometry constraints with two normal-related losses. The direct normal loss $\mathcal{L}_{\text{normal}}$ supervises the predicted normal map $\mathbf{N}_{i,p}$ using angular distance:

$$\mathcal{L}_{\text{normal}} = \sum_{i=1}^{N} \sum_{p \in \mathcal{M}^{\text{valid}}} \arccos\left(\frac{\mathbf{N}_{i,p}^{\top} \hat{\mathbf{N}}_{i,p}}{\|\mathbf{N}_{i,p}\| \|\hat{\mathbf{N}}_{i,p}\|} \right), \tag{15}$$

where $\hat{\mathbf{N}}_{i,p}$ denotes the ground-truth surface normal. We further introduce a geometry-derived normal loss $\mathcal{L}_{\text{points_normal}}$, which follows the same angular formulation as Eq. 15 but replaces the explicit normal prediction with the normal analytically computed from the predicted point map $\mathbf{P}_{i,p}$. This loss encourages the predicted 3D structure to preserve locally coherent surface geometry. The video depth completion teacher is optimized with the same loss formulation as LDCM Yu et al. (2026).

B.2 Implementation Details

ViGeo adopts ViT-G Dosovitskiy et al. (2020) as the backbone. Following recent feed-forward geometry models Lin et al. (2026), one-third of the attention layers are configured with dynamic chunking attention. The backbone features are passed to a 5-layer transformer decoder that applies self-attention within each frame, followed by separate convolutional heads Wang et al. (2025d) for geometric prediction.

Training is conducted in two stages. In the first stage, ViGeo is trained for 50K iterations with a fixed pixel budget of 112,896. In the second stage, it is fine-tuned for 200K iterations with variable resolutions, where the pixel budget is randomly sampled between 112,896 and 268,324. Across both stages, the batch size varies from 2 to 24 samples, and the aspect ratio is randomly sampled from $[0.5, 2.0]$. The backbone is initialized from the pretrained DA3 weights Lin et al. (2026). After the two-stage training, we freeze the preceding network modules and independently optimize the confidence head following Pi3 Wang et al. (2026b).

We use the AdamW optimizer with a cosine learning rate schedule and linear warmup. The peak learning rates are set to 5×10^{-5} and 1×10^{-5} for the first and second stages, respectively, and the backbone learning rate is scaled by 0.1. We apply standard data augmentations, including random cropping, color jittering, Gaussian blur, JPEG compression artifacts, and perspective-aware cropping. The first stage is trained on 16 NVIDIA H20 GPUs, while the second stage uses 48 NVIDIA H20 GPUs. The complete training process takes approximately 20 days.

The attention and refinement ablations in Tables 5 and 6 use a lighter ViT-L backbone trained for 200K iterations. They share the main depth evaluation protocol. The attention ablation uses raw supervision and evaluates each training scheme under offline full-sequence and online frame-wise inference; the weights remain fixed across inference modes, with only the attention mask changing. The refinement ablation compares raw and refined supervision under offline inference with the same DCA configuration.

C Related Work

Dense monocular geometry estimation. Early methods Eigen et al. (2014); Fu et al. (2018); Yin et al. (2019); Bhat et al. (2021); Yuan et al. (2022); Bae et al. (2021); Qi et al. (2020); Yu et al. (2023) are generally restricted to in-domain datasets, severely limiting their generalization to unseen environments. MiDaS Ranftl et al. (2020; 2021); Birkl et al. (2023) pioneers a paradigm shift: by introducing an affine-invariant objective, it unifies diverse data sources for large-scale joint training, drastically improving zero-shot capabilities for relative depth estimation. Building upon this

trajectory, subsequent approaches Yang et al. (2024b;c); Wang et al. (2025d); Yin et al. (2023); Hu et al. (2024a); Piccinelli et al. (2024; 2025); Guizilini et al. (2023) further scale the training corpus, while recent diffusion-based methods Ke et al. (2024); Rombach et al. (2022); Wang et al. (2025b) successfully harness the strong generative priors of latent diffusion models. More recently, drawing inspiration from multi-task learning, joint depth and surface normal estimation methods Fang et al. (2025); Wang et al. (2025e;b); Fu et al. (2024); He et al. (2024); Bae & Davison (2024); Ke et al. (2025) have emerged, effectively leveraging the mutual benefits of these complementary geometric representations. Despite significant progress, by processing images in isolation, existing monocular estimators naturally lack multi-view geometric consistency, leading to severe scale ambiguities and temporal flickering across different viewpoints.

Dense video geometry estimation. Moving beyond isolated frames, dense video geometry estimation fundamentally aims to recover temporally coherent and spatially accurate geometry from video sequences. A common approach is to jointly optimize depth across multiple images using classic and learnable dense visual SLAM methods Teed & Deng (2021); Wang et al. (2024a), or to globally align the outputs of single-image estimators Wang et al. (2023); Luo et al. (2020). Recently, some approaches Lu et al. (2025); Zhang et al. (2025) have also demonstrated that DUSt3R Wang et al. (2024b) can generalize to videos and dynamic scenes. However, a shared bottleneck across all these paradigms is their heavy reliance on computationally expensive post-optimization. Driven by the rapid advancements in feed-forward foundation models Wang et al. (2025a); Gao et al. (2025); Fang et al. (2025); Wang et al. (2025c); Yang et al. (2025); Wang et al. (2026b); Tang et al. (2025); Leroy et al. (2024); Chen et al. (2025); Hu et al. (2024b); Shao et al. (2024); Yang et al. (2024a); Lin et al. (2026), the trend has recently shifted towards directly predicting consistent geometry from video sequences in a purely feed-forward manner. Building upon the robust priors of Depth Anything Yang et al. (2024c), Video Depth Anything Chen et al. (2025) devises an efficient spatiotemporal head and a temporal consistency loss to enforce temporal coherence. Concurrently, DepthCrafter Hu et al. (2024b) unleashes the potential of latent video diffusion models Blattmann et al. (2023) to generate highly consistent, open-world video depth sequences. Furthermore, recent feed-forward 3D reconstruction models Wang et al. (2025a; 2026b); Leroy et al. (2024); Gao et al. (2025) have demonstrated that temporal coherence can also be effectively achieved via alternating attention mechanisms, while concurrently revealing that multi-task learning paradigms (*e.g.*, joint estimation of depth, point maps, and surface normals) significantly enhance overall geometric representation. However, the majority of these architectures are primarily designed for offline inference, where the full sequence is available, and are not naturally suited for streaming or causal settings. To address sequential input scenarios, several recent works have explored streaming 3D reconstruction with causal or persistent memory designs Wang et al. (2025c); Wang & Agapito (2024); Zhuo et al. (2025); LAN et al. (2026); Chou et al. (2025). CUT3R Wang et al. (2025c) introduces a continuous 3D perception model with a persistent state, while FlashDepth Chou et al. (2025) leverages a recurrent network to perform online alignment. More recently, StreamVGGT Zhuo et al. (2025) and Stream3R LAN et al. (2026) extend large-scale geometric transformers to streaming settings through causal architectures. Although these methods improve scalability for long sequences and enable online inference, they typically trade off global context and reconstruction quality compared to offline models. As a result, existing approaches still separate offline and streaming reconstruction into different model designs, leaving a critical gap for a unified framework that can flexibly handle both regimes without retraining.

Large-scale data training. Recently, large-scale data training coupled with advanced network backbones Oquab et al. (2023); Ranftl et al. (2021); Dosovitskiy et al. (2020) has emerged as a powerful paradigm for 3D geometry estimation Wang et al. (2025a; 2026b); Gao et al. (2025); Fang et al. (2025); LAN et al. (2026); Zhuo et al. (2025). Due to the lack of high-quality labeled 3D datasets, various data engines Yang et al. (2024b;c); Wang et al. (2025e); Lin et al. (2026) have been devised. Depth Anything Yang et al. (2024b;c) scales the training datasets by unleashing the power of large-scale unlabeled data, but such paradigms are fundamentally restricted to relative disparity estimation. To obtain more reliable 3D supervision, MoGe2 Wang et al. (2025e) enhances the annotations of noisy datasets Sun et al. (2020) via Poisson reconstruction, while Depth Anything 3 Lin et al. (2026) directly aligns monocular depth maps with sparse measurements. However, these approaches heavily rely on the initial outputs of monocular depth estimation, leaving the final refined annotations inherently bounded by the errors of the underlying monocular estimators. Inspired by the progress of depth completion foundation models Yu et al. (2026); Tan et al. (2026), we devise a data engine based on multi-view depth completion, fully leveraging the strengths from both images and

Table 8: **Quantitative results for surface normal estimation** on the Sintel Bozic et al. (2021), HAMMER Jung et al. (2023), and NYUv2 Silberman et al. (2012) datasets. We report the mean and median angular errors, as well as the threshold accuracy ($\delta_{11.25^\circ} \uparrow$). ViGeo results use the updated model after extended training. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Sintel			HAMMER			NYUv2		
	Mean ↓	Med ↓	$\delta_{11.25^\circ}$ ↑	Mean ↓	Med ↓	$\delta_{11.25^\circ}$ ↑	Mean ↓	Med ↓	$\delta_{11.25^\circ}$ ↑
DSINE Bae & Davison (2024)	40.69	35.52	21.22	12.90	8.36	68.21	16.50	9.17	58.75
NormalCrafter Bin et al. (2025)	**35.39**	**28.57**	24.34	11.11	7.83	70.05	15.87	8.70	60.41
Lotus He et al. (2024)	37.37	29.89	24.03	12.06	9.47	61.53	17.25	9.43	58.00
ViGeo	36.93	28.89	**26.05**	**9.32**	**6.16**	**78.29**	**15.11**	**8.51**	**61.30**

Table 9: **Quantitative results for monocular depth estimation** on the Sintel Bozic et al. (2021), Bonn Palazzolo et al. (2019), and KITTI Geiger et al. (2013) datasets. We report the absolute relative error (Rel) and threshold accuracy (δ_1). The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Sintel		Bonn		KITTI	
	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑
VDA Chen et al. (2025)	0.529	0.407	0.349	0.460	0.256	0.580
DepthCrafter Hu et al. (2024b)	0.291	0.619	0.084	0.928	0.192	0.678
VGGT Wang et al. (2025a)	0.254	0.707	0.047	0.974	0.077	0.929
Pi3 Wang et al. (2026b)	0.272	0.724	**0.044**	**0.975**	0.055	**0.977**
DA3 Lin et al. (2026)	0.281	0.673	0.053	0.963	0.073	0.954
FlashDepth Chou et al. (2025)	0.288	0.664	0.061	0.967	0.084	0.939
ViGeo	**0.217**	**0.745**	0.050	0.969	**0.054**	0.970

sparse measurements to generate dense, accurate, and temporally consistent depth annotations for large-scale training.

D Additional Experimental Results

D.1 Surface Normal Estimation

For surface normal estimation, we evaluate ViGeo on the Sintel Bozic et al. (2021) and HAMMER Jung et al. (2023) video datasets, as well as the NYUv2 Silberman et al. (2012) image benchmark. We compare against the image-based methods DSINE Bae & Davison (2024) and Lotus He et al. (2024), as well as the video-based normal estimator NormalCrafter Bin et al. (2025). Following standard protocols, we report the mean and median angular errors (Mean ↓, Med ↓) and the threshold accuracy within 11.25° ($\delta_{11.25^\circ} \uparrow$). As shown in Table 8, ViGeo achieves the best results across all metrics on HAMMER and NYUv2 and the highest threshold accuracy on Sintel. These results show that ViGeo supports reliable surface normal estimation alongside depth and point map prediction.

D.2 Monocular Depth Estimation

For monocular depth estimation, we evaluate ViGeo on Sintel Bozic et al. (2021), Bonn Palazzolo et al. (2019), and KITTI Geiger et al. (2013). This setting assesses the single-frame geometry estimation capability of ViGeo by applying the model independently to each image, without using temporal context. We compare against representative monocular and feed-forward geometry baselines under the same evaluation protocol. Following standard monocular depth evaluation, we apply affine-

Table 10: **Ablation study on inference schemes for long video depth estimation.** All evaluations are performed on sequences of 300∼400 frames. Peak VRAM is measured on the Bonn Palazzolo et al. (2019) dataset with a per-image token count of 1369. "OOM" indicates out-of-memory on a 96 GB GPU.

Scheme	Bonn		KITTI		HAMMER		VRAM
	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	**(GB)** ↓
Full Sequence	OOM	OOM	OOM	OOM	OOM	OOM	>96.0
Chunked ($C = 16$)	0.059	0.967	0.059	0.972	0.023	0.997	**23.03**
Chunked ($C = 32$)	0.059	0.967	0.057	0.973	0.022	0.997	23.91
Chunked ($C = 48$)	0.059	0.968	0.055	0.975	0.023	0.997	26.86
Chunked ($C = 64$)	0.059	0.967	0.057	0.974	0.023	0.997	29.22

Table 11: **Efficiency comparison.** We report the maximum number of images processed on a 96 GB NVIDIA H20 GPU, model parameters, and average running speed per image. Running speed is measured using 16 images at a resolution of 518×518.

Model	Max # Images	Params.	Speed
VGGT Wang et al. (2025a)	200–250	1.26B	10.96 FPS
ViGeo (ours)	700–800	1.26B	10.00 FPS

invariant alignment, i.e., scale and shift alignment, between the predicted and ground-truth depth for each frame. As shown in Table 9, ViGeo achieves the best performance on Sintel and remains competitive on Bonn and KITTI. These results indicate that, beyond its temporal modeling capability, ViGeo also learns strong single-frame geometric priors and produces reliable depth estimates under the monocular setting.

D.3 Inference Strategy for Long-Video Processing

Benefiting from our proposed dynamic chunking attention design, our model is able to process arbitrarily long video sequences via a KV-cache mechanism Yuan et al. (2026). We ablate different inference schemes by varying the chunk length $C \in \{16, 32, 48, 64\}$. As shown in Table 10, standard full-sequence inference (i.e., without chunking) scales quadratically in memory complexity, inevitably leading to out-of-memory (OOM) errors on long video sequences (∼400 frames). In contrast, our chunking mechanism strictly bounds the memory footprint by limiting the maximum KV-cache size. Although the peak VRAM increases slightly with larger C due to the overhead of intermediate activations, the depth estimation accuracy remains highly consistent across different chunk sizes. Consequently, we adopt $C = 16$ as our default setting to minimize the instantaneous memory peak while maintaining robust temporal consistency for long-range video depth estimation.

D.4 Inference Efficiency

Table 11 compares the full-attention inference capacity and running speed of ViGeo and VGGT Wang et al. (2025a) under the same 96 GB NVIDIA H20 GPU setting. Although both models have the same parameter count, ViGeo supports a larger full-attention input capacity, processing 700–800 images compared with 200–250 images for VGGT. Meanwhile, ViGeo maintains a comparable running speed, achieving 10.00 FPS versus 10.96 FPS on 16 images at 518×518 resolution. It is worth noting that this table reports the maximum number of images under full-attention inference. For long video inference, ViGeo can process arbitrarily long sequences through dynamic chunking attention, which is compatible with KV caching.

D.5 Streaming Inference Latency

The full-sequence efficiency results in Table 11 measure offline inference. To directly evaluate streaming behavior, we follow the evaluation protocol of StreamVGGT Zhuo et al. (2025) and

Table 12: Streaming inference latency in milliseconds (lower is better). Columns indicate the position of the current frame in the sequence. Best results are **bolded**.

Method	Frame 1	Frame 5	Frame 10	Frame 20	Frame 30	Frame 40
ViGeo	**44.81**	**45.25**	**48.66**	**55.28**	**61.51**	**68.03**
StreamVGGT	52.98	54.17	57.42	65.24	71.30	79.30
Stream3R	59.36	71.80	88.17	120.68	153.32	186.00

Table 13: Additional long-video depth comparison using sliding-window inference for VGGT and DA3 and KV-cache inference for Stream3R. Results for the remaining methods are reported in Table 3. Best results are **bolded**, including ties.

	Bonn (400)		KITTI (300)		HAMMER (300)	
Method	Rel↓	δ_1 ↑	Rel↓	δ_1 ↑	Rel↓	δ_1 ↑
VGGT	0.067	0.955	0.128	0.859	0.107	0.863
DA3	0.062	0.958	0.075	0.957	**0.023**	0.994
Stream3R	0.074	0.959	0.101	0.913	0.037	0.984
ViGeo	**0.059**	**0.967**	**0.059**	**0.972**	**0.023**	**0.997**

Stream3R LAN et al. (2026) and measure per-frame latency as the sequence grows. As shown in Table 12, ViGeo has the lowest latency at every reported sequence position. Its latency increases from 44.81 ms at the first frame to 68.03 ms at frame 40, compared with 79.30 ms for StreamVGGT and 186.00 ms for Stream3R at frame 40. These results characterize the response time of frame-wise causal inference as temporal context accumulates.

D.6 Additional Long-Video Comparison

We extend the long-video evaluation in Table 3 by applying overlap-aligned sliding-window inference, following Video Depth Anything Chen et al. (2025), to the offline models VGGT Wang et al. (2025a) and DA3 Lin et al. (2026). For Stream3R LAN et al. (2026), we use the KV-cache strategy of InfiniteVGGT Yuan et al. (2026). Table 13 reports results on 400-frame Bonn sequences and 300-frame KITTI and HAMMER sequences. ViGeo achieves the lowest relative error on Bonn and KITTI, ties DA3 on HAMMER relative error, and obtains the highest δ_1 on all three datasets.

D.7 Adaptation to Varying Chunk Lengths

Depending on sensor bandwidth and latency requirements, video observations may arrive as individual frames, short bursts, or mini-batches. We simulate these conditions on KITTI using randomly varying chunk lengths and sequences of up to 300 frames. Table 14 evaluates a single shared ViGeo checkpoint in two modes: *Stream* disables bidirectional interaction within each chunk, whereas *Chunk* enables it. The inputs and model weights are identical between these two modes. Enabling within-chunk attention reduces Rel from 0.061 to 0.051 and improves δ_1 from 0.970 to 0.977. This comparison shows that DCA can exploit jointly available observations and adapt to different temporal input organizations without retraining or switching checkpoints.

D.8 Post-Training Adaptation of Dynamic Chunking Attention

To test whether dynamic chunking attention (DCA) transfers beyond ViGeo, we replace the global-attention modules of VGGT-Ω Wang et al. (2026a) and DA3 Lin et al. (2026) with causal attention or DCA. All remaining components are frozen, and only the replacement attention modules are optimized for 50K iterations. We evaluate the original global-attention models and the DCA variants offline, and the causal-attention and DCA variants online. Table 15 reports depth estimation on Bonn and camera pose estimation on TUM-Dynamics. DCA improves offline depth accuracy while largely preserving pose performance. Under online inference, DCA improves all reported metrics over the corresponding causal variants for both models, supporting its use as a lightweight post-training strategy for adding streaming capability to existing geometry models.

Table 14: KITTI evaluation under streaming inputs with randomly varying chunk lengths. Both ViGeo modes use the same checkpoint. Best results are **bolded**.

Method	AbsRel↓	δ_1 ↑
Video Depth Anything Chen et al. (2025)	0.377	0.431
DepthCrafter Hu et al. (2024b)	0.321	0.454
GeometryCrafter Xu et al. (2025b)	0.083	0.961
VGGT Wang et al. (2025a)	0.183	0.699
Pi3 Wang et al. (2026b)	0.110	0.862
DA3 Lin et al. (2026)	0.082	0.932
Stream3R LAN et al. (2026)	0.101	0.913
StreamVGGT Zhuo et al. (2025)	0.207	0.658
FlashDepth Chou et al. (2025)	0.100	0.919
ViGeo (Stream mode)	0.061	0.970
ViGeo (Chunk mode)	**0.051**	**0.977**

Table 15: Post-training adaptation of dynamic chunking attention (DCA) to VGGT-Ω and DA3. Only the replacement attention modules are trained for 50K iterations. Best results within each model and inference setting are **bolded**, including ties.

			Bonn		TUM-Dynamics		
Model	Attn.	Mode	Rel↓	δ_1 ↑	ATE↓	RPE-t↓	RPE-r↓
VGGT-Ω	Global	Offline	0.043	0.971	**0.007**	**0.278**	**0.006**
VGGT-Ω	DCA	Offline	**0.038**	**0.980**	**0.007**	0.279	**0.006**
VGGT-Ω	Causal	Online	0.052	0.980	0.013	0.287	0.008
VGGT-Ω	DCA	Online	**0.048**	**0.981**	**0.011**	**0.282**	**0.007**
DA3	Global	Offline	0.050	0.972	0.012	**0.309**	0.009
DA3	DCA	Offline	**0.048**	**0.976**	**0.010**	0.310	**0.008**
DA3	Causal	Online	0.055	0.971	0.016	0.332	0.011
DA3	DCA	Online	**0.052**	**0.974**	**0.012**	**0.323**	**0.009**

D.9 Additional Data-Refinement Ablations

We first analyze how different stages of the refinement pipeline affect label quality on 50 ScanNet++ Yeshwanth et al. (2023) scenes. Table 16 evaluates both frame-wise depth quality and temporal consistency as the raw measurements pass through Poisson reconstruction, outlier filtering, and depth completion. To isolate the effect of the completion method, LDCM and the video completion teacher are applied as alternative branches to the same filtered and Poisson-reconstructed input. Filtering improves the temporal quality of the reconstructed labels, while the video completion teacher provides the strongest overall accuracy and consistency.

We then train two otherwise identical ViGeo models using either raw or refined labels and evaluate them on HAMMER Jung et al. (2023), keeping the architecture and training protocol fixed. Compared with the main data-refinement ablation in Table 6, Table 17 additionally reports TAE. The model trained with refined labels improves both frame-wise depth accuracy and temporal consistency, showing that the gains in label quality transfer to the trained model rather than remaining specific to the refinement pipeline.

D.10 Additional Attention-Pattern Ablation

The main attention ablation in Table 5 focuses on depth accuracy across standard video benchmarks. Table 18 complements it by reporting TAE for full, causal, and dynamic chunking attention under both offline and online inference. Dynamic chunking achieves the lowest TAE in both settings while retaining strong depth accuracy, demonstrating that its flexible temporal context also improves the stability of video predictions.

Table 16: Label-level ablation of the data-refinement pipeline on 50 ScanNet++ scenes. We evaluate the depth accuracy and temporal consistency (TAE) of labels produced by each pipeline variant. LDCM and the video completion teacher are alternative completion strategies applied to the same filtered and Poisson-reconstructed input.

Base processing	Completion method	Rel↓	RMSE↓	δ_1 ↑	TAE↓
Raw	–	0.031	0.138	0.975	3.594
Poisson	–	0.031	0.134	0.975	3.535
Filter → Poisson	–	0.031	0.133	0.975	3.314
	LDCM	0.031	0.137	0.975	3.570
	Video Completion Teacher	**0.026**	**0.117**	**0.984**	**2.578**

Table 17: Effect of data refinement on model training. ViGeo models trained with raw or refined labels are evaluated on HAMMER using 90-frame inputs for depth accuracy and temporal consistency (TAE).

Training supervision	Rel↓	δ_1 ↑	TAE↓
Raw labels	0.033	0.995	0.262
Refined labels (Ours)	**0.028**	**0.996**	**0.255**

Table 18: Attention-pattern ablation on HAMMER.

Attention scheme	Rel↓	TAE↓
Offline		
Full Attention	0.034	0.270
Causal Attention	0.035	0.273
Dynamic Chunking (Ours)	**0.033**	**0.262**
Online		
Full Attention	0.037	0.322
Causal Attention	0.036	0.313
Dynamic Chunking (Ours)	**0.034**	**0.301**

E Additional Visualizations

E.1 Additional Point Cloud Visualizations

Fig. 7 presents additional qualitative results of point cloud reconstruction. ViGeo produces accurate and realistic 3D structures with coherent global geometry and fine local details. The reconstructed point clouds preserve object boundaries and scene layouts, demonstrating the effectiveness of ViGeo in recovering dense and spatially consistent geometry from videos.

E.2 Additional Qualitative Results

Fig. 8 compares monocular depth estimation results from a single input image. Our method produces sharp object boundaries and consistent relative geometry. Compared with VGGT and Pi3, our results better preserve clean foreground–background separation with fewer blurred or noisy structures. Although DepthCrafter yields sharp object contours, its background depth is overly smooth and its relative geometry is often inaccurate, leading to incorrect depth ordering and distorted scene structure. In contrast, our method maintains both crisp boundaries and geometrically plausible depth.

Fig. 9 compares video depth estimation results across consecutive frames. Beyond per-frame depth quality, our method maintains stable depth structures over time, especially for moving objects and camera-induced scene changes. VGGT and Pi3 often produce noisy responses and noticeable frame-to-frame depth fluctuations, causing unstable object shapes and inconsistent background geometry. DepthCrafter generates smoother results, but it tends to over-smooth large background regions



Figure 7: Additional point cloud visualizations. ViGeo produces accurate and realistic reconstructions with coherent geometry and fine structural details.

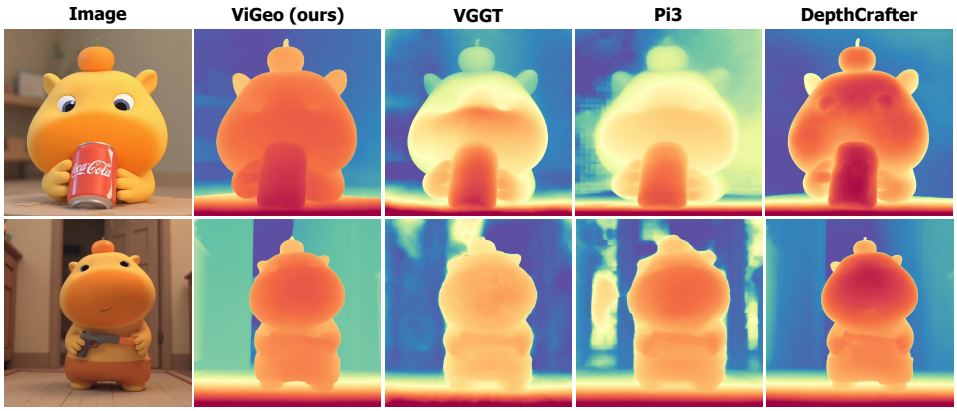

Figure 8: Qualitative results for monocular depth estimation. Compared with existing methods, ViGeo produces sharper depth boundaries and more accurate relative geometry.

and exhibits temporal inconsistency around dynamic objects. In contrast, our method preserves coherent depth ordering and stable object/background structures across frames, producing sharper and temporally consistent video depth predictions.

E.3 Additional Visualizations of Completion-Based Data Refinement

Fig. 10 presents additional qualitative results of completion-based data refinement. Compared with raw depth maps that suffer from large missing regions, our pipeline produces dense and geometrically consistent labels. The refined point clouds further recover complete surfaces and sharp structures that are fragmented or absent in the raw sensor measurements.

F Limitations

While ViGeo achieves promising results, high-resolution and 4D video geometry estimation remain challenging. High-resolution inputs may introduce additional computational cost, particularly for long sequences. Moreover, more explicit 4D representations could be explored to further improve temporal consistency in dynamic scenes. Despite these remaining challenges, we believe ViGeo provides a useful step toward scalable video geometry understanding.

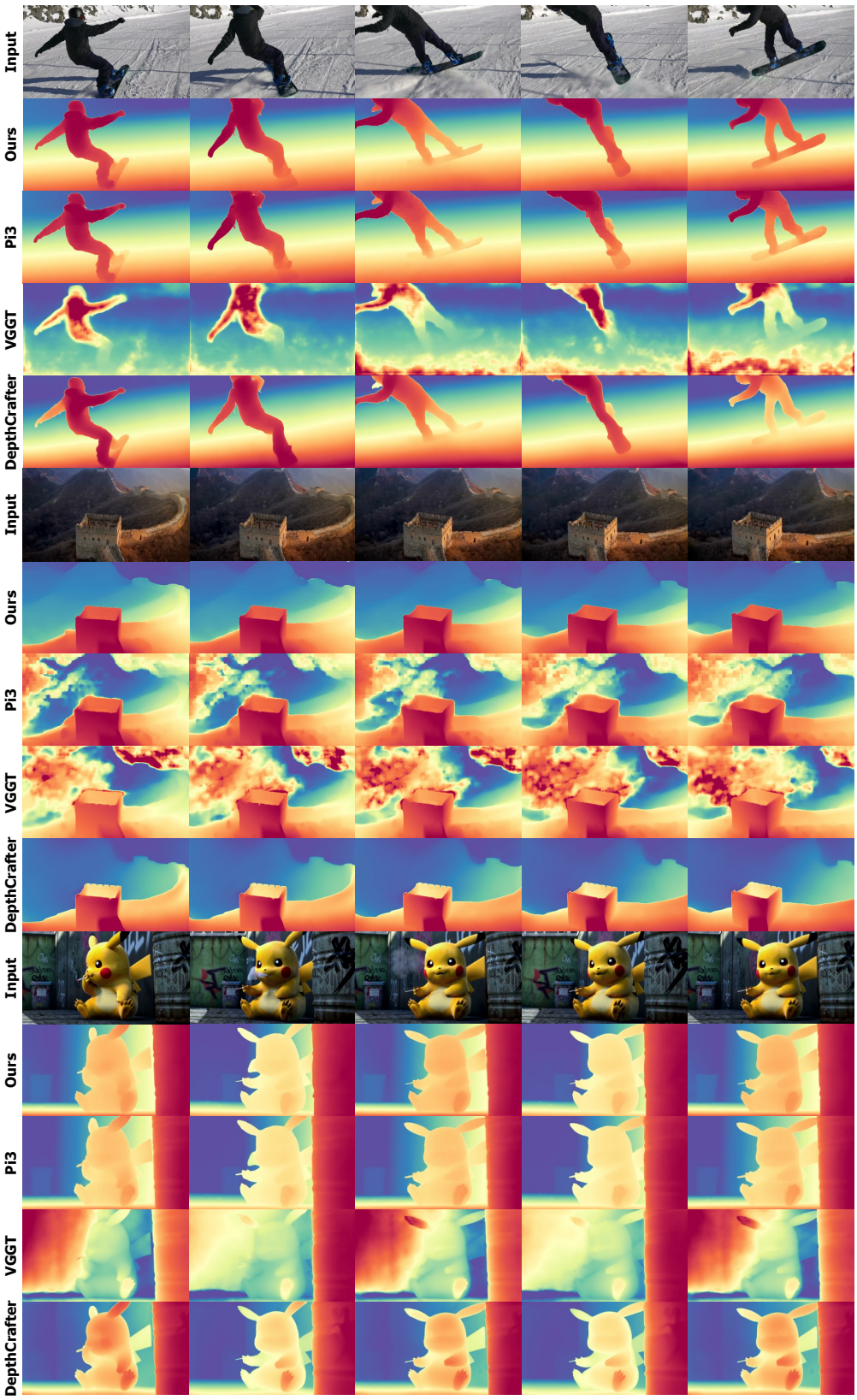

Figure 9: Qualitative results for video depth estimation. Compared with existing methods, ViGeo produces sharper, more accurate, and temporally stable depth.

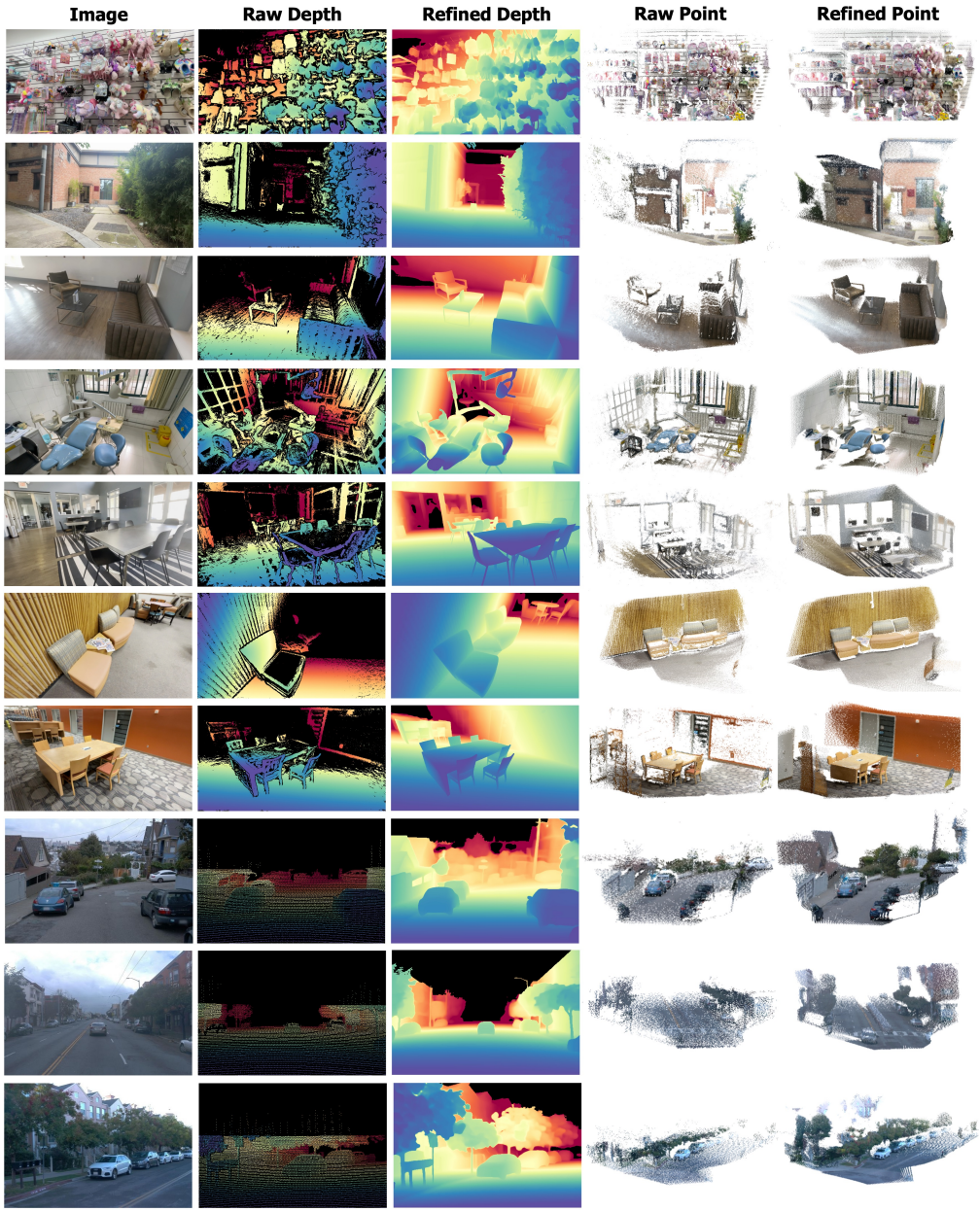

Figure 10: Qualitative results of the data refinement pipeline. We visualize the sparse raw measurements (cols. 2 & 4) alongside our refined pseudo-labels (cols. 3 & 5). While the raw depth suffers from extensive missing regions, our pipeline produces spatially dense and geometrically consistent labels. Notably, the refined point clouds recover solid surfaces and sharp structural details that are fragmented or entirely absent in the raw sensor outputs.