Cambrian-*P*: Pose-Grounded Video Understanding

Jihan Yang[1,*], Zifan Zhao[1,*], Xichen Pan[1], Shusheng Yang[1], Junyi Zhang[2], Hu Xu[3], Shang-Wen Li[3], and Saining Xie[1]

[1] New York University
[2] UC Berkeley
[3] Meta FAIR
* Equal contribution.

Abstract. Camera pose matters. The position and orientation of each viewpoint define a shared spatial coordinate frame that relates observations across video frames. Yet this signal is absent from multimodal LLMs (MLLMs) for video understanding, which process frames as isolated 2D snapshots, instead of the persistent scene humans perceive. We revisit pose as a lightweight supervisory signal and introduce Cambrian-*P*, a video MLLM augmented with per-frame learnable camera tokens and a pose regression head. With a carefully designed sampling scheme, the model achieves substantial gains of 4.5–6.5% on spatial reasoning benchmarks such as VSI-Bench, generalizes across eight additional spatial and general video QA benchmarks, and, as a byproduct, achieves state of the art streaming pose estimation on ScanNet. Surprisingly, training on pseudo-annotated poses from in-the-wild video further improves general video QA benchmarks, showing pose helps beyond spatial reasoning. Together, these results position camera pose as a fundamental signal for video models that reason about the physical world.

Keywords: Video Understanding, Camera Pose Estimation

1 Introduction

Video is the 2D observation of a dynamic 3D scene from a coherent sequence of viewpoints. Each viewpoint is defined by the observer's pose, *i.e.*, its 3D position ($\mathbb{R}^3$) and orientation ($SO(3)$), specifying how the camera is embedded in the physical world [25,51]. Pose provides the link between pixels and geometry, serving as a global anchor that relates distinct views to a shared coordinate frame [25, 67]. With pose, a video is no longer a collection of ambiguous image projections, but a unified and coherent 3D scene [67, 73].

Recovering camera poses from visual observations has therefore been a cornerstone of 3D vision and robotics for decades. Structure from Motion (SfM) emerged in the 1980s precisely for this purpose [49, 67], and remains a prerequisite for a wide array of downstream tasks: multi-view stereo [69], 3D reconstruction, and neural rendering with NeRF [54] or 3D Gaussian Splatting [35]. In parallel, robotics and augmented-reality systems depend on SLAM [13] and

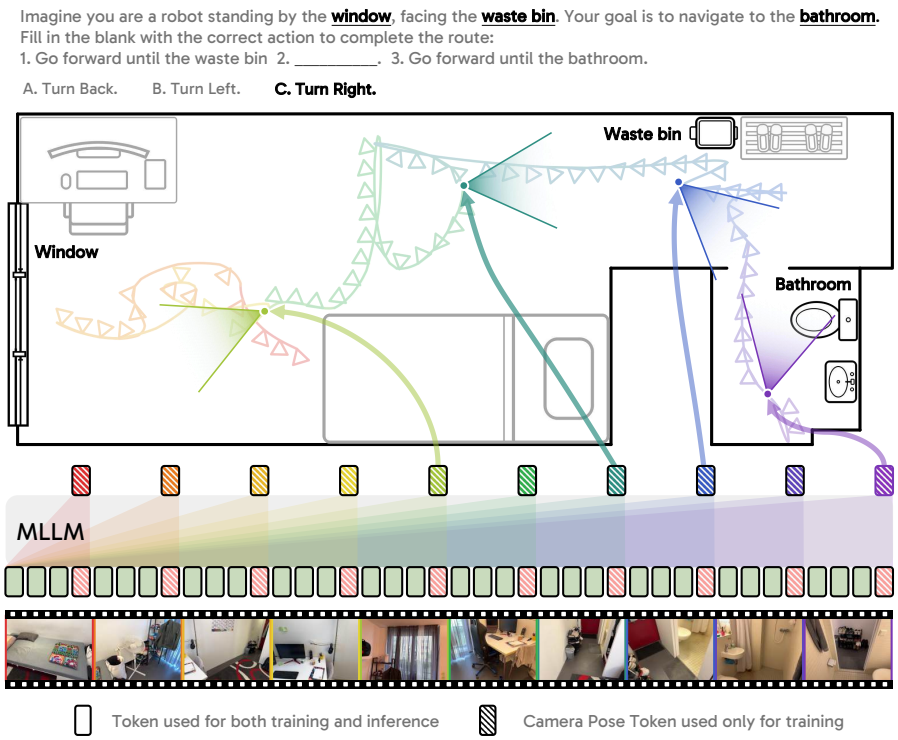


Fig. 1: Cambrian-*P* illustration in video QA. Cambrian-P equips the current video understanding paradigm with native camera pose predictions by an extra camera token per frame. Cambrian-P positions video frames into a shared spatial coordinate frame, then effectively models the underlying 3D world projected in video.

visual odometry [57] to localize themselves and reason about their own motion. More recently, the community has recognized that jointly predicting pose and depth with feed-forward transformers can recover dense 3D structure in a single pass [43,84], further underscoring the centrality of pose as a geometric primitive.

Yet, the role of pose has remained confined to 3D vision. We argue that its value extends far beyond. Multimodal large language models (MLLMs) [1,5,11, 24, 47, 77, 79, 80] now excel at semantic video understanding—recognizing actions, summarizing narratives, and answering questions—but consistently struggle when tasks demand spatial reasoning [36, 76, 93, 94]. We contend that this failure is not incidental. Without explicit grounding in 3D geometry, each frame is processed as an independent 2D snapshot, disconnected from the spatial structure across views. Our key insight is that pose is uniquely suited to close this gap. It is the lightest 3D signal, compactly encoding how views relate geometrically; it enforces global consistency through rigid-body (SE(3)) constraints; and it disentangles camera motion from scene dynamics, collapsing the space of plausible spatial interpretations. These properties make pose not a useful auxiliary cue, but a foundational inductive bias for spatially aware video understanding.

In this work, we present Cambrian-P, which grounds pixels with camera poses by introducing pose supervision into MLLMs, calling for a new paradigm for video understanding. Cambrian-P introduces minimal architectural overhead: a single learnable camera token is appended to each frame's visual features, and a lightweight projector and head regress pose parameters from the LLM's hidden states. However, naively adding a pose regression loss does not work out of the box. Through careful investigation, we find that pose estimation and video question answering rely on fundamentally different frame sampling and data augmentation strategies. Uniform temporal sampling common in video understanding provides a shortcut for memorizing poses, while the heavy augmentation typical in pose estimation distracts from semantic comprehension. To reconcile these tensions, we design an interleaved training strategy that incorporates pose-only training data processed and augmented following its preferred practices, along with a random-jitter frame sampling strategy that introduces controlled perturbation to the conventional uniform sampling used for VQA.

With this lightweight design and optimized joint training, Cambrian-P achieves state-of-the-art results in both spatial VQA and streaming camera pose estimation. On VSI-Bench [94] and VS*Temporal*I-Bench [22], Cambrian-P obtains 4.5%∼6.5% gains over its no-pose counterpart, and demonstrates strong out-of-distribution generalization across eight spatial and general VQA benchmarks, including MindCube [102], MVBench [40], and EgoSchema [52]. Scaling pose supervision to in-the-wild videos with pseudo annotations further improves general video QA benchmarks, positioning pose as a promising signal for video understanding beyond spatial reasoning. For pose estimation, Cambrian-P delivers state-of-the-art streaming camera pose accuracy on ScanNet ATE, outperforming specialist reconstruction models such as StreamVGGT [113], CUT3R [86], and Point3R [90]. We further observe consistent scalability with respect to model size, data size, and training iterations for both tasks. Finally, our analysis reveals two insights: (i) while adding depth supervision may seem intuitive for injecting 3D priors, it proves suboptimal for VQA compared to learning from camera pose; and (ii) camera pose enables MLLMs to think more globally across video frames, evidenced by the significant improvement of Cambrian-P when answering questions about spatially distant objects than nearby ones.

2 Related Work

Multimodal Large Language Models Driven by the tremendous success of Large Language Models (LLMs) [1,5,11,24,79,80] in linguistic understanding and reasoning, alongside powerful pretrained visual representations [26,58,62,81,105], Multimodal Large Language Models (MLLMs) [6, 37, 39] extend LLMs beyond language-only corpora and have achieved impressive progress in understanding visual media such as images [16, 18, 37, 42, 46, 74, 77] and videos [7, 64, 71, 85, 96, 109]. However, despite their remarkable success in semantic parsing [2,17], world knowledge acquisition [27, 66, 103], and general reasoning [50, 104], MLLMs are still far from achieving human-level embodied intelligence capable of perceiving, reasoning, and acting within the 3D real world [36, 76, 93]. Recent studies [63,

94,100,102] pinpoint that a fundamental deficit hindering existing MLLMs from this goal is their unsatisfactory visual spatial intelligence, which serves as one of the foundational elements for humans to understand the 3D outside world but remains largely absent in modern MLLMs. This paper aims to bridge this gap.

Visual Spatial Intelligence The growing interest in grounding MLLMs in the real 3D world has created an urgent need to improve their visual spatial intelligence: the ability to understand underlying spatial geometry from visual inputs. Motivated by this, recent works have proposed various benchmarks to evaluate this capability using single-image [63], multi-image [102], or video inputs [94] (which is the primary focus of our work). Their results suggest that even frontier MLLMs still fall significantly behind human performance in spatial understanding. To bridge this gap, several studies [10,15,22,95,96] curate spatial-oriented data by repurposing existing 3D-related datasets [4,9,20,65,101], applying pseudo-labeling, or designing synthetic data generation pipelines [21]. These efforts not only improve models' spatial understanding but also provide foundational datasets for future exploration. [48, 59, 95] propose to finetune MLLMs on spatial data using reinforcement learning to improve their spatial reasoning capability. Another line of research [22, 38, 111] introduces 3D features from off-the-shelf 3D encoders [84]. While this significantly improves MLLMs' spatial awareness, the approach remains inflexible as it is largely constrained by the quality of the pre-trained features. Recent work [28] unifies 3D reconstruction with spatial understanding with a dual-encoder and mixture-of-transformers design, which is heavy and yields suboptimal results.

Camera Pose Estimation Camera pose estimation serves as a foundational pillar of 3D vision. It is not merely an isolated task, but the essential prerequisite for a wide spectrum of downstream applications, ranging from dense multi-view reconstruction [23, 67, 98] to modern neural rendering [35, 54] and robotic navigation [13, 61]. Traditionally, recovering camera extrinsics relies on classic SfM and SLAM systems [25, 55, 67]. While mathematically elegant, these heuristic-based pipelines frequently struggle in ill-posed scenarios characterized by textureless regions, repetitive patterns, or highly dynamic environments. Recently, a paradigm shift toward data-driven, feed-forward 3D estimation has emerged, with methods like DUSt3R [87] and MASt3R [56] bypassing fragile heuristics via direct dense pointmap regression, giving rise to a broad family of follow-up works. Among them, offline models [34, 43, 84, 89] jointly process multiple views and typically offer stronger bidirectional reasoning over the full observation set, while streaming approaches [83, 86, 113] process frames incrementally, making them better suited for arbitrary-length videos.

In this work, we contextualize the data-driven learning of 3D geometry within the broader paradigm of MLLM spatial reasoning. Rather than relying on specialized vision architectures or heavy dual-encoder designs, we highlight the camera pose as a lightweight signal that connects isolated frames into a continuous 3D space. By unifying continuous camera pose estimation and video understanding within a single MLLM, our proposed Cambrian-P not only yields competitive

streaming pose estimation but fundamentally endows the MLLM with a coherent, global understanding of the 3D physical world.

3 Cambrian-P

We introduce Cambrian-P, a new video understanding paradigm for multimodal large language models by equipping it with native camera pose estimation capability. We start by introducing our framework in Sec. 3.1, followed by the training objective and dynamics in Sec. 3.2 and Sec. 3.3, respectively.

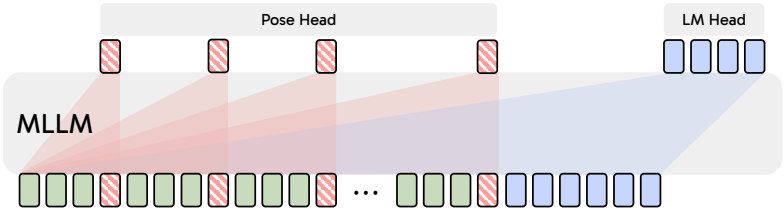


Fig. 2: Cambrian-P architecture illustration. Cambrian-P imposes minimal modifications to current MLLM architectures, introducing only two learnable pose tokens and a lightweight pose head. These tokens are appended to visual tokens, while positioned before text embeddings.

3.1 Architecture

As illustrated in Fig. 2, our overall architecture introduces minimal additional components and overhead during both training and inference to enable camera pose estimation in the current MLLMs paradigm.

MLLM. We build Cambrian-P upon the Cambrian-S [96] architecture, a pre-trained MLLM that pairs a SigLIP2-SO400m [81] vision encoder with a Qwen2.5 [91] language model connected via an MLP projector.

Camera Pose Tokens. To enable camera pose estimation within the LLM's feature space, we introduce a small set of learnable camera pose tokens that are appended to each frame's visual tokens before they enter the LLM, inspired by the practice of VGGT [84]. Specifically, we define two learnable queries $\mathbf{c}_{\text{first}}, \mathbf{c}_{\text{rest}} \in \mathbb{R}^H$, where H is the LLM's hidden dimension. For a sequence of N frames, we assign $\mathbf{c}_{\text{first}}$ to the first frame and $\mathbf{c}_{\text{rest}}$ to all remaining frames. This allows the model to distinguish the first frame from the rest, and to represent all poses in the coordinate system of the first camera. The per-frame token sequence fed to the LLM is:

$$[\,\mathbf{v}_i^{(1)}, \ldots, \mathbf{v}_i^{(K)}\,;\,\mathbf{c}_i\,], \quad i = 1, \ldots, N, \tag{1}$$

where $\mathbf{v}_i^{(j)}$ denotes the K projected visual tokens, and $\mathbf{c}_i = \mathbf{c}_{\text{first}}$ for $i = 1$, $\mathbf{c}_i = \mathbf{c}_{\text{rest}}$ for $i > 1$. Note that $\mathbf{c}_i$ is placed after the vision tokens of each frame due to the causal attention mechanism of the LLM. After the LLM forwarding, we extract and slice out the pose token hidden state $\mathbf{h}_i \in \mathbb{R}^H$ for each frame from its final layer hidden states.

Camera Pose Projector and Head. We bridge the LLM and the camera prediction head with a linear camera pose projector that maps LLM's hidden representation $\mathbf{h}_i$ to the required camera pose feature dimension as $\tilde{\mathbf{h}}_i = \mathbf{W}_p \mathbf{h}_i$. To regress the camera parameters for each frame from $\{\tilde{\mathbf{h}}_i\}$, we adopt the camera pose head design of VGGT [84], which consists of four self-attention layers followed by a linear prediction layer.

3.2 Training Objective

Our training objective combines the next-token prediction loss for vision-language understanding with a camera pose estimation loss for spatial understanding. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{NTP}} + \lambda_{\text{pose}} \cdot \mathcal{L}_{\text{pose}}, \tag{2}$$

where $\mathcal{L}_{\text{NTP}}$ is the standard cross-entropy loss over response text tokens, $\mathcal{L}_{\text{pose}}$ is the camera pose estimation loss, and λ_{pose} is a weighting coefficient.

Camera Pose Estimation Loss. Following VGGT [84], we represent each camera as a pose encoding $\mathbf{g}_i = [\mathbf{t}_i, \mathbf{q}_i, \mathbf{f}_i] \in \mathbb{R}^9$, where $\mathbf{t}_i \in \mathbb{R}^3$ is the translation, $\mathbf{q}_i \in \mathbb{R}^4$ is the rotation quaternion, and $\mathbf{f}_i = [f_i^h, f_i^w] \in \mathbb{R}^2$ encodes the horizontal and vertical field of view. The camera pose loss supervises predicted pose encodings $\hat{\mathbf{g}}_i$ against ground-truth encodings $\mathbf{g}_i$ using a weighted ℓ_1 loss:

$$\mathcal{L}_{\text{pose}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\frac{w_T}{\bar{d}} \|s^* \hat{\mathbf{t}}_i - \mathbf{t}_i\|_1 + w_R \|\hat{\mathbf{q}}_i - \mathbf{q}_i\|_1 + w_f \|\hat{\mathbf{f}}_i - \mathbf{f}_i\|_1 \Big), \tag{3}$$

where w_T, w_R, and w_f denote the component weights, and $\bar{d}$ is the normalization factor for trajectory scale.

As training data can span a wide range of physical scales from indoor scenes to large outdoor driving sequences, the magnitude of translation errors can vary by orders of magnitude. Furthermore, non-metric datasets inherently possess arbitrary numerical scales, which would otherwise lead to unpredictable gradient magnitudes. To prevent large-scale scenes or arbitrarily scaled non-metric data from dominating the gradient, we normalize the translation loss term by the sequence-averaged consecutive frame distance of the ground-truth trajectory, defined as $\bar{d} = \frac{1}{N-1} \sum_{i=2}^{N} \|\mathbf{t}_i - \mathbf{t}_{i-1}\|_2$, which ensures that both scene types contribute comparable gradients during training.

To train on both metric and non-metric datasets [41,45], we resolve the scale ambiguity of non-metric data before computing the translation loss. For non-metric samples, we align predicted translations to ground truth using a least-squares scale factor $s^* = \text{stop-grad}\Big(\frac{\sum_i \hat{\mathbf{t}}_i \cdot \mathbf{t}_i}{\sum_i \hat{\mathbf{t}}_i \cdot \hat{\mathbf{t}}_i}\Big)$, treating s^* as constant during backpropagation to avoid trivial solutions. For metric samples, $s^* = 1$.

3.3 Improving Training Dynamics

While the architecture and training objective of Cambrian-P are straightforward, the training dynamics present the most significant challenge when jointly optimizing VQA and camera pose estimation.

Fig. 3: Interleaved training of Cambrian-*P*. Top: augmented pose-only samples using dynamic frame sampling and only pose supervision. Bottom: samples using uniform frame sampling with both VQA and pose supervision. L is the total number of frames of the video.

Training Dynamic Gaps between VQA and Camera Pose Estimation. Challenges primarily come from three conflicts in their training paradigms. *First*, a video frame sampling gap exists: MLLMs typically sample frames at fixed intervals regardless of the query. This yields repeated ground-truth poses across iterations, encouraging memorization of video-pose correspondences rather than genuine pose learning. In contrast, robust camera pose estimation requires random starting frames and dynamic temporal intervals in frame sampling [34, 84, 86]. *Second*, there is a gap in training duration. Advanced MLLMs typically train for only a single epoch, whereas pose estimation models require tens of epochs with diverse frame sampling to converge [84, 86]. *Third*, the data augmentation gap complicates joint training. While VQA training generally omits augmentations to preserve the factual correctness of answers, camera pose estimation relies on augmentations like color jittering, Gaussian blur, and grayscale [84]. We empirically observe that applying these data augmentations to pose estimation samples is crucial for pose estimation and simultaneously benefits VQA performance.

Interleaved Training between VQA and Pose. To overcome the aforementioned gaps, we introduce an interleaved training strategy that incorporates dedicated pose estimation samples using their preferred sampling and augmentation strategy with only the pose loss. Specifically, given $\hat{M}$ training samples with pose supervision, we augment them by a ratio of β. The resulting $\lfloor \beta \hat{M} \rfloor$ augmented samples follow the standard sampling and augmentation strategies used in camera pose estimation and are trained with only the camera pose loss $\mathcal{L}_{\text{pose}}$. We omit the VQA loss here, as the limited temporal coverage of these samples lacks sufficient context for question answering (see Fig. 3). Applying the VQA objective on incomplete visual information could encourage hallucination. Furthermore, the augmentation of pose-only samples enables us to arbitrarily scale training iterations for pose estimation, fully decoupled from the VQA objective. As shown in Fig. 3, in our implementation, our batches are fully mixed, including samples with VQA-only, pose-only, or joint supervision.

Random Jitter Frame Sampling. In addition, we apply a random jitter augmentation to the uniformly sampled frame indices in VQA. Specifically, given a set of uniformly sampled frame indices $\{u_i\}$ from a video of N total frames, we perturb each index u_i by a random offset $\delta_i \sim \mathcal{U}(-\Delta, \Delta)$, where $\Delta = \lfloor N \cdot \alpha \rfloor$

and α is a jitter ratio controlling the perturbation magnitude. To maintain sequence validity, the jittered index is then clipped to $[0, u_{i+1} - 1]$ for intermediate frames and $[0, N - 1]$ for the last frame, and monotonicity is enforced to ensure $u_i \geq u_{i-1}$ after perturbation. This simple strategy introduces controlled temporal variability to VQA sampling, alleviating memorization of fixed frame-pose correspondences in the uniform frame sampling.

Implementation Details. We finetune Cambrian-P from Cambrian-S-7B stage 3 [96] following its stage 4 training recipe. We perform end-to-end finetuning with AdamW optimizer with learning rates of 1×10^{-5} for the LLM and vision projector, 2×10^{-6} for the vision encoder, and 1×10^{-4} for the pose projector and head, respectively. The pose projector and head are randomly initialized and trained from scratch. By default, we set the interleaved training augmentation ratio β to 1, the random jitter ratio α to 0.005, and the loss trade-off factor λ_{pose} is empirically set to 0.2. We train Cambrian-P on 64 H200 GPUs with a 256 batch size throughout. For training data, we use VSI-590K [96] and data from MapAnything [34]. When only partial labels are available, Cambrian-P activates only the corresponding supervision loss, *i.e.*, VQA loss or camera pose loss.

4 Improved VQA with Cambrian-P

4.1 Experiment Setups

Training Setups. For fair comparison with Cambrian-S and existing MLLMs, we train Cambrian-P with only data from VSI-590K unless otherwise specified.

Benchmarks. We evaluate on a comprehensive suite of spatial reasoning and video understanding benchmarks, including VSI-Bench [94], VSTIBench [22], SparBench [106], MMSIBench [97], MMSIVideo [44], MindCube [102], Tomato [70], MVBench [40], EgoSchema [52], and Perception Test [60].

Baselines. We compare Cambrian-P against three categories of models: (1) general-purpose MLLMs including GPT-4o [32], Gemini-2.5 Pro [19], Qwen2.5VL-7B [8], InternVL-3/3.5 [112], and Qwen3-VL [7]; (2) spatial-specialist models including VST [95], VLM-3R [22], VG-LLM [111], SenseNoVA-SI [14], Cambrian-S [96], and GeoThinker [38]; and (3) chance-level baselines (random and frequency).

4.2 Results

Main Results. As shown in Tab. 1, Cambrian-P yields state-of-the-art spatial reasoning capability in VSI-Bench. In particular, compared to existing spatial-specialist models with the same LM like Cambrian-S-7B, SenseNova-SI-8B, and GeoThinker-7B, Cambrian-P achieves more than a 5% gain. Moreover, Cambrian-P outperforms Cambrian-S †, its counterpart without camera pose estimation, by 4.5%, highlighting the effectiveness of incorporating camera pose prediction in spatial reasoning. For per-subtask improvement, Cambrian-P shows the most prominent improvement on absolute distance, relative direction, and route plan – tasks that demand a more global understanding of the space. It is also noteworthy that Cambrian-P shows superior out-of-distribution generalization capability in the route plan task, which is not included in the VSI-590K training set.

Table 1: VSI-Bench Results Comparison. † indicates this Cambrian-S is fine-tuned only on VSI-590K.

Model	LM		Numerical Answer				Multiple-Choice Answer			
		Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Baselines										
Chance Level (Random)	–	–	–	–	–	–	25.0	36.1	28.3	25.0
Chance Level (Frequency)	–	34.0	62.1	32.0	29.9	33.1	25.1	47.9	28.4	25.2
General-purpose Models										
GPT-4o [32]	Unk.	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-2.5 Pro [19]	Unk.	51.5	43.8	34.9	64.3	42.8	61.1	47.8	45.9	71.3
Qwen2.5VL-7B [8]	Qwen2.5-7B	29.3	25.2	10.5	36.4	29.6	38.4	38.0	29.8	26.8
InternVL-3 8B [112]	Qwen2.5-7B	42.1	68.1	39.0	48.4	33.6	48.3	36.4	27.3	35.4
InternVL-3.5 8B [112]	Qwen3-8B	56.3	–	–	–	–	–	–	–	–
Qwen3-VL 8B [7]	Qwen3-8B	56.6	–	–	–	–	–	–	–	–
Spatial-specialist Models										
VST 7B [95]	Qwen2.5-7B	61.2	71.6	43.8	75.5	69.2	60.0	55.6	44.3	69.2
VLM-3R 7B [22]	Qwen2-7B	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
VG-LLM 8B [111]	Qwen2.5-7B	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2
Cambrian-S 7B [96]	Qwen2.5-7B	67.5	73.2	50.5	74.9	72.2	71.1	76.2	41.8	80.1
SenseNova-SI 8B [14]	Qwen2.5-7B	68.7	–	–	–	–	–	–	–	–
GeoThinker 7B [38]	Qwen2.5-7B	68.5	–	–	–	–	–	–	–	–
GeoThinker 8B [38]	Qwen3-8B	72.6	–	–	–	–	–	–	–	–
Cambrian-S-7B† [96]	Qwen2.5-7B	69.2	73.6	53.7	75.2	74.7	71.5	82.0	38.7	84.3
Cambrian-P	Qwen2.5-7B	**73.7**	**74.9**	**60.1**	**76.0**	**76.9**	**74.8**	**89.5**	**52.6**	**85.0**

Table 2: VS*Temporal*I-Bench Result. We finetune Cambrian-P on VSI-590K and VLM-3R [22] data for this experiment. Cambrian-P shows 20% improvement on the camera movement direction subtask.

Methods	Avg.	Cam-Obj Abs. Dist.	Cam. Displace.	Cam. Mov. Dir.	Obj-Obj Rel. Pos.	Cam-Obj Rel. Dist.
GPT-4o [32]	38.2	29.5	23.4	37.3	58.1	42.5
Gemini-1.5 Flash [75]	32.1	28.5	20.9	24.4	52.6	33.9
LLaVA-NeXT-Video-72B [47]	44.0	32.3	10.5	48.1	78.3	50.9
VLM-3R-7B [22]	58.8	39.4	39.6	60.6	86.5	68.6
GeoThinker 8B [38]	67.4	38.4	45.8	84.2	93.6	**75.2**
Cambrian-P (w/o Pose)	62.4	39.4	40.6	67.7	92.2	72.0
Cambrian-P	**68.9**	**42.5**	**46.6**	**87.7**	**94.3**	73.2

Significant improvement in Understanding Camera Movement. To further evaluate how well Cambrian-P captures camera movement, we finetune it on VSI-590K [96] and VLM-3R [22] data and evaluate on VSTI-Bench [22], which includes questions about camera motion. As shown in Tab. 2, Cambrian-P achieves state-of-the-art results on VSTI-Bench. More importantly, it obtains a 20% improvement on the camera movement subtask over the no-pose baseline, demonstrating that the camera pose estimation objective directly enhances the model's understanding of camera dynamics.

Table 3: Out-of-Distribution Generalization for Spatial and General VQA Benchmarks. Cambrian-P is fine-tuned only on VSI-590K, without any in-distribution training data for benchmarks here.

Model	SparBench	MMSIBench	MMSIVideo	MindCube	MVBench	EgoSchema	Perception Test	Tomato
Cambrian-P (w/o Pose)	32.7	26.2	20.1	34.3	51.9	49.6	56.4	25.1
Cambrian-P	**35.9**	**28.0**	**22.9**	**38.4**	**53.5**	**52.5**	**58.4**	**26.7**

Improvement on OOD Spatial and General VQA Benchmarks. As shown in Tab. 3, although Cambrian-P is finetuned solely on VSI-590K, which is in-distribution with respect to VSI-Bench, it also demonstrates improvements on out-of-distribution spatial and general VQA benchmarks. This suggests that

the local-to-global video understanding capability acquired through camera pose prediction is a general and fundamental skill transferable to broader VQA tasks.

4.3 Improving General Video QA with Pseudo-Annotated Pose

Cambrian-P shows promising improvements on both spatial VQA and general VQA benchmarks when ground-truth pose supervision is available. However, GT camera poses are available only for limited data sources in VSI-590K (*e.g.*, ScanNet, ScanNet++, and ARKitScenes). To scale pose supervision to general-domain videos, we pseudo-annotate videos corresponding to the subsampled 590K samples from Cambrian-S-3M [96]. We curate pseudo poses using VIPE [30]. Video clips first pass a scene-cut detector and a quality filter based on Qwen3-VL [7]. Remaining clips are processed by VIPE and post-filtered, and the resulting pseudo poses are used as GT poses in the interleaved training recipe.

As shown in Tab. 4, adding general VQA data substantially improves general video QA performance on MVBench, Perception Test, and EgoSchema, but slightly degrades VSI-Bench. Introducing GT pose supervision reverses this spatial degradation, improving VSI-Bench while preserving the gains on general VQA. Adding pseudo-pose supervision from general-domain videos further boosts all four benchmarks, yielding additional gains on VSI-Bench as well as substantial improvements on MVBench and EgoSchema. These results suggest that pseudo poses, even when derived from noisy in-the-wild videos, provide a scalable supervision signal for video understanding.

Table 4: Cambrian-P with general VQA training data and pseudo pose supervision. We report Cambrian-P 128 frames results on VSIBench, MVBench, Perception Test, and EgoSchema.

Training Data	Pose Sup.	% Pose Sup.	VSIBench	MVBench	Perception Test	EgoSchema
Spatial VQA Data Only						
VSI-590K	–	0%	71.2	51.7	56.7	48.5
VSI-590K	GT	49%	73.7	53.8	58.1	51.3
Spatial VQA + General VQA Data						
VSI-590K + CamS-590K	–	0%	70.9	68.0	66.9	71.2
VSI-590K + CamS-590K	GT	25%	73.7	67.9	67.8	71.7
VSI-590K + CamS-590K	GT + Pseudo	48%	**73.9**	**69.3**	**67.9**	**73.6**

5 Camera Pose Estimation with Cambrian-P

5.1 Experiment Setups

Training Setups. To further push the camera pose estimation capability, we train Cambrian-P on data with pose annotation from VSI-590K (*i.e.*, ScanNet [20], ScanNet++ [101], and ARKitScenes [9]) and datasets from MapAnything [34], which include metric-scale datasets (ParallelDomain4D [82], TartanAir-v2 [88,110], MVS-Synth [31], Spring [53], SailVOS3D [29], ETH3D [68], Dynamic Replica [33], MPSD [3], and UnrealStereo4K [78]) and non-metric-scale datasets (MegaDepth [41], DL3DV [45], and BlendedMVS [99]). To further boost the performance on camera pose estimation, we set the interleaved training augmentation ratio β to 20 and the loss trade-off factor λ_{pose} to 0.5.

Table 5: Camera pose estimation results on ScanNet, TUM, and Sintel.

Model	ScanNet			TUM-dynamic			Sintel		
	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓	ATE ↓	RPE trans ↓	RPE rot ↓
Offline Models									
VGGT [84]	0.035	0.015	0.380	**0.009**	**0.008**	0.350	0.172	0.061	0.470
DUSt3R-GA [87]	0.081	0.028	0.784	0.083	0.017	3.567	0.417	0.250	5.796
MASt3R-GA [56]	0.078	0.020	0.475	0.038	0.012	0.448	0.185	0.060	1.496
MonST3R-GA [107]	0.077	0.018	0.529	0.098	0.019	0.935	0.111	0.044	0.869
Fast3R [92]	0.155	0.123	3.491	0.090	0.101	1.425	0.371	0.298	13.750
FLARE [108]	0.064	0.023	0.971	0.026	0.013	0.475	0.207	0.090	3.015
π^3 [89]	**0.031**	**0.013**	**0.347**	0.014	0.009	**0.312**	**0.074**	**0.040**	**0.282**
MapAnything [34]	0.052	0.025	0.720	0.029	0.023	0.370	0.226	0.077	0.640
Streaming Models									
StreamVGGT [113]	0.127	0.041	1.880	0.062	0.030	0.690	0.273	0.109	0.850
CUT3R [86]	0.096	**0.022**	**0.590**	**0.045**	**0.015**	**0.440**	**0.215**	**0.070**	**0.630**
Point3R [90]	0.097	0.035	2.791	0.058	0.031	0.758	0.442	0.154	1.897
Spann3R [83]	0.096	0.023	0.661	0.056	0.021	0.591	0.329	0.110	4.471
G^{2}VLM [28]	0.148	0.048	1.220	0.129	0.044	0.700	0.301	0.135	1.450
Cambrian-P	**0.078**	0.023	0.880	0.046	0.020	0.580	0.239	0.081	2.440

Benchmarks. We evaluate camera pose estimation on three standard benchmarks: ScanNet [20], TUM-dynamic [72], and Sintel [12], covering indoor scenes, handheld sequences, and synthetic movies with large camera motions. Following MonST3R [107], for TUM-dynamic and ScanNet, we sample the first 90 frames with a temporal stride of 3, and for Sintel, we exclude static scenes or sequences with near-straight camera motion. We report three metrics: Absolute Trajectory Error (ATE), Relative Pose Error in translation (RPE trans), and Relative Pose Error in rotation (RPE rot). All metrics are computed with Sim(3) alignment.

Baselines. We compare Cambrian-P against two categories of methods. Offline methods that require access to all frames simultaneously include VGGT [84], DUSt3R [87], MASt3R [56], MonST3R [107], Fast3R [92], FLARE [108], π^3 [89], and MapAnything [34], all evaluated with global alignment (GA) where applicable. Streaming methods that process frames incrementally include StreamVGGT [113], CUT3R [86], Point3R [90], Spann3R [83], and G^{2}VLM [28].

5.2 Results

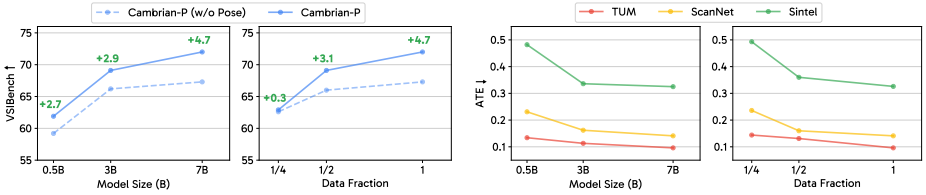
As shown in Tab. 5, Cambrian-P achieves the minimal ATE on ScanNet [20] among all streaming camera pose estimation models and delivers competitive performance on TUM [72], and Sintel [12], without relying on specialized designs like DINOv2 encoder [58] or bidirectional transformer [84, 86]. This highlights that standard MLLMs can predict accurate camera pose with only an additional pose head and two learnable pose queries. In addition, harvesting the compact representation from SigLIP encoder [81], lower FLOPs of causal transformer, and optimized inference infrastructure of LLM community, Cambrian-P shows competitive latency despite its large model size (refer to Appendix).

6 Analysis

To better understand the property of Cambrian-P, we delve into the details of it with extensive experiments. Without other specifications, all Cambrian-P variants are trained with VSI-590K in 32 frames and 196 tokens per frame.

6.1 Scaling Model, Data and Training

The remarkable success of LLMs and the next-token prediction paradigm can be largely attributed to their scalability. We investigate whether the camera pose estimation objective exhibits similar scaling behavior within the MLLM paradigm, across model size, data size, and training iterations.



(a) VQA acc. with various models and data sizes. **(b)** Pose ATE with various models and data sizes.

Fig. 4: Comparison of Cambrian-*P* regarding different model size and data size. Larger models or more data both yield higher VSI-Bench scores and lower pose estimation error across all benchmarks.

Model Size Scaling. To investigate the effect of model size, we finetune Cambrian-*P* from Cambrian-*S* variants with different LLM sizes. As shown in Fig. 4a (a), scaling up the model not only improves VSI-Bench performance but also widens the gap over the no-pose baseline. We attribute this trend to the inherent demands of multi-task learning, where larger model capacity better accommodates the additional complexity. Also, as shown in Fig. 4b (a), the ATE of camera pose consistently decreases as the model size increases.

Data Size Scaling. As shown in Fig. 4a (b), scaling up data size similarly improves VSI-Bench performance while widening the gap over the no-pose baseline. Note that Cambrian-*P* yields only marginal improvement with $\frac{1}{4}$ data, likely because the pose head is trained from scratch and struggles to converge with limited supervision. We empirically find that pretraining the pose head beforehand can alleviate this issue. Also, as illustrated in Fig. 4b (b), the translation error of the camera pose also consistently decreases with larger data size.

Training Iteration Scaling. As shown in Tab. 6, scaling augmented pose iterations improves VQA performance more efficiently than increasing VQA iterations. Even without extra interleaved pose iterations, pose supervision yields a 2% gain. Meanwhile, Tab. 7 shows that more pose iterations consistently reduce ATE, demonstrating the scalability of MLLMs for pose estimation. Notably, even with large-scale out-of-distribution pose data and a dominant pose objective, Cambrian-*P* still improves VSI-Bench performance, highlighting the synergy between spatial VQA and camera pose estimation.

6.2 Ablation Studies

Component Ablation As shown in Tab. 8, pose loss with interleaved training improves VSI-Bench by about 3% and reduces pose ATE. Random jitter adds another 0.8% VQA gain with better pose accuracy, confirming that both designs mitigate the VQA-pose training gap.

Table 6: Scaling training iterations to improve VQA.

Model	VQA Iteration	Pose Iteration	VSI-Bench ↑	ScanNet ATE↓	TUM ATE ↓	Sintel ATE ↓
Cambrian-P (w/o Pose)	N	0	67.3	-	-	-
	2N	0	69.3	-	-	-
	3N	0	69.3	-	-	-
Cambrian-P	N	0	69.4	0.259	0.132	0.521
	0	N	25.2	0.163	0.115	0.374
	N	$\frac{N}{2}$	72.0	**0.141**	**0.096**	**0.325**
	N	N	72.2	0.149	0.112	0.329
	2N	N	**72.7**	0.143	0.106	0.406

Table 7: Scaling pose iterations with MapAnything data.

Model	VQA Iteration	Pose Iteration	VSI-Bench ↑	ScanNet ATE↓	TUM ATE ↓	Sintel ATE ↓
Cambrian-P	N	0	67.3	-	-	-
	N	N	**71.6**	0.145	0.105	0.361
	N	3N	70.9	0.106	0.073	0.297
	N	5N	69.8	0.094	0.071	0.289
	N	20N	69.3	**0.077**	**0.048**	**0.278**

Number of Frames Ablation. As shown in Tab. 9, Cambrian-P improves on VSI-Bench as the number of input frames increases, while its margin over the baseline shrinks. This suggests that extra frames provide richer visual context for QA, whereas pose estimation favors lower inter-frame overlap and is often trained on shorter sequences of 12∼24 frames [84]. Consistently, ATE on pose benchmarks slightly degrades with more frames.

6.3 How Does Camera Pose Help Video QA?

Here, to understand how camera pose supervision helps, we investigate individual loss components and analyze VQA accuracy across varying spatial distances.

Camera Pose Helps MLLMs Learn. Starting from a standard MLLM, Cambrian-P introduces pose tokens for pose supervision during training and conditions on them at inference. We ask whether Video QA gains come from inference-time pose trajectories or from better representations learned through pose supervision. As shown in Tab. 10, training with pose tokens and supervision yields significant gains across video benchmarks, while inference-time pose conditioning adds no benefit. This suggests that Cambrian-P improves by learning stronger pose-supervised representations, rather than by using pose at inference.

Camera Pose Helps VQA More than Depth. As shown in Tab. 11, adding pose loss alone improves VSI-Bench accuracy by 2.6% over depth loss. Combining both losses leads to a slight degradation in both VQA and camera pose estimation over the pose loss alone baseline. This indicates that camera pose estimation has greater synergy with video understanding as probed by VQA. Breaking down the pose loss into its components, we find that both translation and rotation losses effectively improve VQA performance, while field-of-view loss yields gains comparable to those from depth loss.

Table 8: Components ablation study of Cambrian-*P*.

Camera Loss	Interleaved Training	Random Jitter	VSI-Bench ↑	ScanNet ATE↓	TUM ATE ↓	Sintel ATE ↓
			67.3	-	-	-
✓	✓		71.2	0.144	0.106	0.366
✓		✓	69.4	0.259	0.132	0.521
✓	✓	✓	**72.0**	**0.141**	**0.096**	**0.325**

Table 9: Ablation on the number of input frames during training.

Model	# frames / # tok	VSI-Bench↑	ScanNet ATE ↓	TUM ATE ↓	Sintel ATE ↓
Cambrian-*P* (w/o Pose)	32 / 196	67.3	–	–	–
Cambrian-*P*		72.0 (+4.7)	0.141	0.096	0.325
Cambrian-*P* (w/o Pose)	64 / 64	70.3	–	–	–
Cambrian-*P*		73.1 (+2.8)	0.140	0.104	0.272
Cambrian-*P* (w/o Pose)	128 / 64	71.2	–	–	–
Cambrian-*P*		73.7 (+2.5)	0.141	0.111	0.322

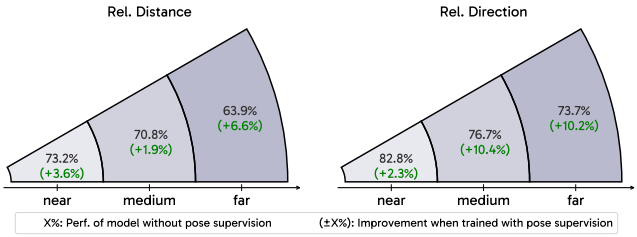


Fig. 5: Camera pose improves global spatial reasoning. Pose gains are larger for far objects: 6.6% vs. 3.6% on relative distance and 10.2% vs. 2.3% on relative direction.

Camera Pose Makes MLLM Thinking in Space More Globally. VSI-Bench [94] observes that MLLMs fall short in spatial intelligence as they tend to see locally rather than globally. Here, we investigate whether enabling the MLLM to be aware of camera movement facilitates more global spatial reasoning. As shown in Fig. 5, we group samples based on normalized ground-truth distance relative to room size into near, medium, and far categories for the relative distance and relative direction question types in VSI-Bench. We find that without pose supervision, model performance degrades as objects become farther apart, while Cambrian-*P* exhibits larger gains for distant objects compared to nearby ones. This indicates that camera pose supervision enables MLLMs to develop more global spatial reasoning capabilities.

6.4 Can Video QA Help Camera Pose Estimation?

We have extensively discussed how the 3D prior from camera pose estimation benefits video QA. But does the reverse also hold—can VQA improve camera pose estimation? As shown in Tab. 12, when the pretrained MLLM is more grounded in video QA in terms of better VSI-Bench [94], EgoSchema [52], Perception Test [60], and MVBench [40] performance, the model finetuned on MapAnything [34] data predicts more accurate camera poses after fine-tuning. We attribute this to the better video-language alignment via VQA pretraining, which provides a more effective foundation for the post-LLM camera pose head.

Table 10: Effect of pose tokens during training and inference in Cambrian-*P*. Cambrian-P's improvements are driven by pose supervision during training, not by pose-token conditioning at inference.

Pose Token Training	Pose Token Inference	VSI-Bench	VSTIBench	MVBench	EgoSchema	Perception Test	Tomato
✗	✗	67.3	55.1	51.9	49.6	56.4	20.4
✓	✓	**72.0**	56.5	**53.5**	**52.5**	58.4	**26.7**
✓	✗	**72.0**	**56.6**	53.2	**52.5**	**58.8**	**26.7**

Table 11: Loss ablation study results. T, R, and FV indicate translation, rotation, and field-of-view loss, respectively.

Pose Loss	Depth Loss	VSI-Bench ↑	ScanNet ATE↓	TUM ATE ↓	Sintel ATE ↓
✗	✗	67.3	-	-	-
✓	✗	**72.0**	**0.141**	0.096	0.325
✓	✓	71.7	0.156	0.118	**0.318**
✗	✓	69.4	0.371	0.194	0.537
T only	✗	70.7	0.205	0.128	0.357
R only	✗	69.7	0.287	**0.092**	0.353
FV only	✗	69.4	0.408	0.196	0.670
T + R	✗	71.5	0.165	0.130	0.386

Table 12: Cambrian-*P* results finetuned from different Cambrian-*S* variants.

Model	VSI-Bench ↑	EgoSchema ↑	Percept. Test ↑	MVBench ↑	ScanNet ATE ↓	TUM ATE ↓	Sintel ATE ↓
CamS-S1	21.4	42.9	44.4	43.9	-	-	-
Cambrian-P (FT CamS-S1)	68.1	-	-	-	0.130	0.085	0.366
CamS-S2	24.6	47.5	53.5	49.2	-	-	-
Cambrian-P (FT CamS-S2)	69.6	-	-	-	0.105	0.073	0.285
CamS-S3	35.7	76.9	70.8	66.3	-	-	-
Cambrian-P (FT CamS-S3)	**69.8**	-	-	-	**0.094**	**0.071**	**0.289**

7 Conclusion

We introduce Cambrian-P, a pose-grounded video understanding model that equips standard MLLMs with the capability to connect individual frames in a shared space. With a simple yet scalable architectural design and tailored training dynamics, Cambrian-P improves spatial and general video QA, and achieves competitive streaming pose estimation performance against state-of-the-art methods. Our results position camera pose as an important missing signal for video MLLMs: it grounds frames in a globally consistent 3D space and encourages learning cross-frame correspondences. Cambrian-P advances MLLMs toward real-world grounded video understanding.

Acknowledgments

We thank Oscar Michel, Baiqiao Yin, Jianyuan Wang, Anjali Gupta, Ellis Brown, Peter Tong, and Pinzhi Huang for reviewing this manuscript and providing constructive feedback. This work is supported by a grant from the Meta FAIR team. S.X. acknowledges support from the MSIT IITP grant (RS-2024-00457882) and the NSF award IIS-2443404.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: ICCV (2019)
3. Antequera, M.L., Gargallo, P., Hofinger, M., Bulo, S.R., Kuang, Y., Kontschieder, P.: Mapillary planet-scale depth dataset. In: ECCV (2020)
4. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: CVPR (2016)
5. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
6. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
7. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
8. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
9. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., Shulman, E.: ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In: NeurIPS (2021)
10. Brown, E., Ray, A., Krishna, R., Girshick, R., Fergus, R., Xie, S.: SIMS-V: Simulated instruction-tuning for spatial video understanding. arXiv preprint arXiv:2511.04668 (2025)
11. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
12. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV (2012)
13. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. T-RO (2017)
14. Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., et al.: Scaling spatial intelligence with multimodal foundation models. arXiv preprint arXiv:2511.13719 (2025)
15. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: CVPR (2024)

16. Chen, G., Li, Z., Wang, S., Jiang, J., Liu, Y., Lu, L., Huang, D.A., Byeon, W., Le, M., Rintamaki, T., et al.: Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In: NeurIPS (2025)
17. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
18. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR (2024)
19. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
20. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
21. Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., Mottaghi, R.: Procthor: Large-scale embodied ai using procedural generation. In: NeurIPS (2022)
22. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In: CVPR (2026)
23. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. TPAMI (2009)
24. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
25. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
26. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
27. Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., Liu, Z.: Video-MMMU: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826 (2025)
28. Hu, W., Lin, J., Long, Y., Ran, Y., Jiang, L., Wang, Y., Zhu, C., Xu, R., Wang, T., Pang, J.: G^2vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. arXiv preprint arXiv:2511.21688 (2025)
29. Hu, Y.T., Wang, J., Yeh, R.A., Schwing, A.G.: Sail-vos 3d: A synthetic dataset and baselines for object detection and 3d mesh reconstruction from video data. In: CVPR (2021)
30. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., et al.: Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934 (2025)
31. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: CVPR (2018)
32. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
33. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Dynamicstereo: Consistent dynamic depth from stereo videos. In: CVPR (2023)

34. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
35. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. (2023)
36. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
37. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. TMLR (2025)
38. Li, H., Cao, Q., Tang, T., Xiang, K., Guo, Z., Han, J., Xu, H., Liang, X.: Thinking with geometry: Active geometry integration for spatial reasoning. arXiv preprint arXiv:2602.06037 (2026)
39. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
40. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: MVbench: A comprehensive multi-modal video understanding benchmark. In: CVPR (2024)
41. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: CVPR (2018)
42. Li, Z., Chen, G., Liu, S., Wang, S., VS, V., Ji, Y., Lan, S., Zhang, H., Zhao, Y., Radhakrishnan, S., et al.: Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. arXiv preprint arXiv:2501.14818 (2025)
43. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
44. Lin, J., Xu, R., Zhu, S., Yang, S., Cao, P., Ran, Y., Hu, M., Zhu, C., Xie, Y., Long, Y., et al.: Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. arXiv preprint arXiv:2512.10863 (2025)
45. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: CVPR (2024)
46. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
47. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
48. Liu, Y., Zhang, B., Zang, Y., Cao, Y., Xing, L., Dong, X., Duan, H., Lin, D., Wang, J.: Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. arXiv preprint arXiv:2510.27606 (2025)
49. Longuet-Higgins, H.C.: A computer algorithm for reconstructing a scene from two projections. Nature (1981)
50. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In: ICLR (2024)
51. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S.: An Invitation to 3-D Vision: From Images to Geometric Models. Springer Science & Business Media (2005)
52. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: NeurIPS (2023)
53. Mehl, L., Schmalfuss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: CVPR (2023)

54. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM (2021)
55. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: A versatile and accurate monocular slam system. T-RO (2015)
56. Murai, R., Dexheimer, E., Davison, A.J.: Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In: CVPR (2025)
57. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: CVPR (2004)
58. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
59. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)
60. Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. In: NeurIPS (2023)
61. Qin, T., Li, P., Shen, S.: Vins-mono: A robust and versatile monocular visual-inertial state estimator. T-RO (2018)
62. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
63. Ramakrishnan, S.K., Wijmans, E., Kraehenbuehl, P., Koltun, V.: Does spatial cognition emerge in frontier models? In: ICLR (2025)
64. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR (2024)
65. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021)
66. Saikh, T., Ghosal, T., Mittal, A., Ekbal, A., Bhattacharyya, P.: Scienceqa: A novel resource for question answering on scholarly articles. IJDL (2022)
67. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
68. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
69. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR (2006)
70. Shangguan, Z., Li, C., Ding, Y., Zheng, Y., Zhao, Y., Fitzgerald, T., Cohan, A.: Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In: ICLR (2025)
71. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR (2024)
72. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: IROS (2012)
73. Szeliski, R.: Computer vision: algorithms and applications. Springer Nature (2022)
74. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)

75. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
76. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
77. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., Pan, X., Wang, Z., Fergus, R., LeCun, Y., Xie, S.: Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In: NeurIPS (2024)
78. Tosi, F., Liao, Y., Schmitt, C., Geiger, A.: Smd-nets: Stereo mixture density networks. In: CVPR (2021)
79. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
80. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
81. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
82. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: ECCV (2024)
83. Wang, H., Agapito, L.: 3d reconstruction with spatial memory. In: 3DV (2025)
84. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
85. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
86. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: CVPR (2025)
87. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
88. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: IROS (2020)
89. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. In: ICLR (2026), https://openreview.net/forum?id=DTQIjngDta
90. Wu, Y., Zheng, W., Zhou, J., Lu, J.: Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. arXiv preprint arXiv:2507.02863 (2025)
91. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
92. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: CVPR (2025)
93. Yang, J., Ding, R., Brown, E., Qi, X., Xie, S.: V-IRL: Grounding virtual intelligence in real life. In: ECCV (2024)

94. Yang, J., Yang, S., Gupta, A., Han, R., Fei-Fei, L., Xie, S.: Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. In: CVPR (2024)
95. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025)
96. Yang, S., Yang, J., Huang, P., II, E.L.B., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Fergus, R., LeCun, Y., Fei-Fei, L., Xie, S.: Towards spatial supersensing in video. In: ICLR (2026)
97. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al.: Mmsi-bench: A benchmark for multi-image spatial intelligence. arXiv preprint arXiv:2505.23764 (2025)
98. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: ECCV (2018)
99. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: CVPR (2020)
100. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. arXiv preprint arXiv:2504.15280 (2025)
101. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023)
102. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: ICLR (2026)
103. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR (2024)
104. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al.: Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In: ACL (2025)
105. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
106. Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. arXiv preprint arXiv:2503.22976 (2025)
107. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: MonST3r: A simple approach for estimating geometry in the presence of motion. In: ICLR (2025), https://openreview.net/forum?id=lJpqxFgWCM
108. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In: CVPR (2025)
109. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. TMLR (2025)
110. Zhang, Y., Keetha, N., Lyu, C., Jhamb, B., Chen, Y., Qiu, Y., Karhade, J., Jha, S., Hu, Y., Ramanan, D., et al.: Ufm: A simple path towards unified dense correspondence with flow. arXiv preprint arXiv:2506.09278 (2025)
111. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3d world: Enhancing MLLMs with 3d vision geometry priors. In: NeurIPS (2025), https://openreview.net/forum?id=oUYmk8WaGO

112. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
113. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming visual geometry transformer. In: ICLR (2026), https://openreview.net/forum?id=5APgTKsnx8