EVGT: EMBODIED VISUAL GEOMETRY TRANSFORMER

Anonymous authors
Paper under double-blind review

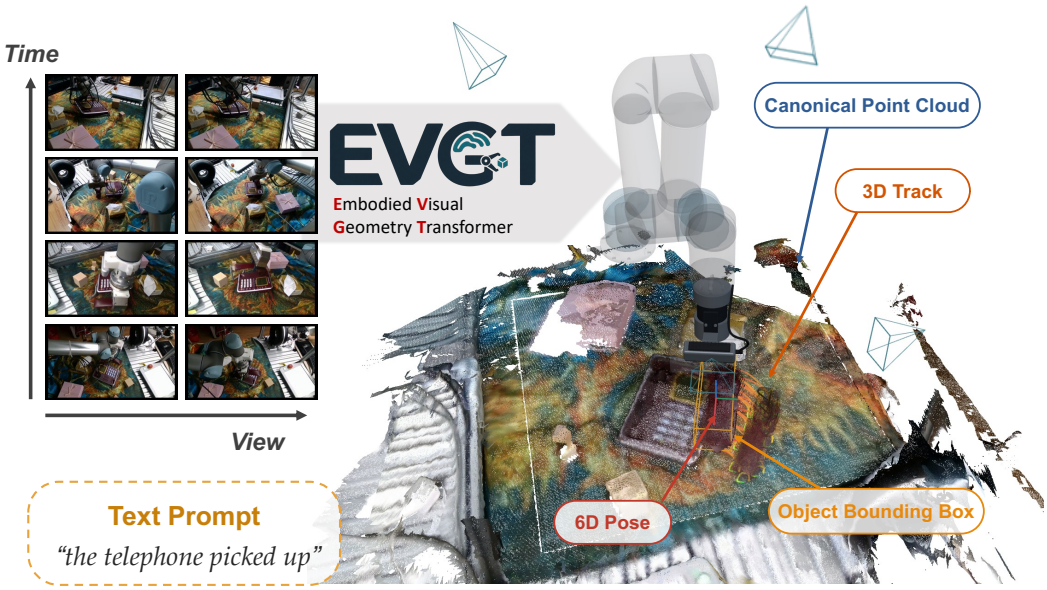


Figure 1: **Embodied Visual Geometry Perception.** Given multi-view RGB sequences with object prompt, EVGT jointly infers canonical scene geometry and object dynamics in a single forward pass. The motion and extent of the object remain consistently registered as camera viewpoints and object configurations evolve, yielding a unified representation for downstream manipulation policies.

ABSTRACT

3D reconstruction is fundamental to robotic perception, providing the spatial structure needed to understand scenes and object motion. Despite recent progress in visual geometry foundation models, a unified geometry foundation model tailored to robotic scenes remains lacking. To address this gap, we propose an **E**mbodied **V**isual **G**eometry **T**ransformer (**EVGT**), a metric-scale geometry foundation model for robot video that integrates multi-view observations to recover scene geometry with the motion and 6D pose of the object of interest. From RGB observations, EVGT predicts camera poses, metric depth, point maps, and dense 3D tracks together with 6D pose, size, oriented box, mask, and visibility of the object across time in a shared reference frame. We extract visual tokens from RGB observations and augment them with object queries, with optional prompt tokens for target specification. A spatiotemporal transformer models interactions among these tokens across views and frames, and multiple prediction heads decode the resulting features into scene geometry, 3D motion, and object states. We adopt a two-stage training strategy that combines heterogeneous scene-geometry supervision with partially labeled object-state data. We report point-map, depth, and camera-pose results on six robot-operation sources, dynamic 3D tracking on four core sources, and object-center and rotation diagnostics on six object sources. In downstream closed-loop manipulation, $\pi_{0.5}$ augmented with frozen EVGT features reaches 86.31% success on LIBERO-Plus and 59.17% on RoboCasa24, compared with 84.67% for the LIBERO-Plus $\pi_{0.5}$ starting policy and 55.1% for a RoboCasa-fine-tuned RGB-only $\pi_{0.5}$ baseline.

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.

1 INTRODUCTION

Rapid advances in robotics are extending artificial intelligence from the digital world into the physical world (Brohan et al., 2023; Kim et al., 2024). Operating in the 3D world, the physical laws governing motion and interaction have a natural and compact expression in 3D space (Murray et al., 1994; Lynch & Park, 2017). 3D reconstruction and understanding are therefore fundamental to both robot navigation and manipulation (Hornung et al., 2013; Gervet et al., 2023; Goyal et al., 2023).

Recent visual geometry foundation models reconstruct cameras, depth, point maps, and tracks from uncalibrated images in a feed-forward manner (Wang et al., 2024; 2025a; Keetha et al., 2025; Wang et al., 2025c). Despite promising results in general (Wu et al., 2025; Zhuo et al., 2025; Wang et al., 2025a;c) and driving scenarios (Zuo et al., 2025; 2026), robotic perception introduces additional geometric challenges. Robotic observations often combine external cameras with cameras mounted on moving wrists, such that the relative poses between multiple views vary over time (Fang et al., 2024). Integrating these observations requires modeling camera geometry across both views and time. Robotic interaction also centers on individual objects, making their geometry, 6D pose, and motion essential components of scene understanding (Gervet et al., 2023; Wen et al., 2024). The representation must therefore preserve object geometry and track object state as both the cameras and objects move. Object-pose systems address part of this problem, but often rely on instance masks, object models, or reference images (Wen et al., 2024). Dynamic geometry models recover scene motion and point trajectories (Zhang et al., 2025b; Sucar et al., 2025; Karhade et al., 2025), yet typically do not expose the pose, extent, visibility, and mask of an object as a unified state across time. However, embodied scenarios still lack a unified geometry foundation model that integrates multi-view observations with temporally varying relative camera poses and jointly reconstructs scene geometry, object 6D pose, and object motion in a common metric reference frame.

To address this gap, we propose an **E**mbodied **V**isual **G**eometry **T**ransformer (**EVGT**), a metric-scale geometry foundation model tailored to embodied scenes, as illustrated in Figure 1. From multi-view RGB sequences, EVGT jointly predicts camera poses, metric depth, point maps, and dense 3D tracks, together with object 6D pose, size, oriented boxes, masks, and visibility over time. These predictions share a common metric reference frame, relating object pose and motion to the surrounding scene as camera viewpoints change. Specifically, we encode RGB observations into visual tokens and introduce object queries to capture object geometry and motion, with optional prompt tokens for target specification. A spatiotemporal transformer integrates these representations across views and frames, aggregating complementary observations under changing relative camera poses and object motion. Multiple prediction heads then decode the resulting features into scene geometry, 3D motion, and object states. We adopt a two-stage training strategy that first learns scene geometry and dense motion from heterogeneous video data, then jointly optimizes scene and object predictions using the annotations available in each dataset. This strategy combines broad geometric supervision with partially labeled object data to learn a shared representation of scenes and objects. We evaluate EVGT on robot video spanning single- and two-camera observations, with metrics covering scene reconstruction, camera trajectories, dense motion, and object-state estimation. We also test its frozen features in $\pi_{0.5}$ policies on LIBERO-Plus and RoboCasa24.

2 RELATED WORK

Geometry Models for Robotic Scenes. Geometric representations for robotics encompass spatial feature fields, object-centric representations, and explicit pose estimation, reflecting the importance of local geometry and object structure for physical interaction (Gervet et al., 2023; Goyal et al., 2023; Zhu et al., 2023; Wen et al., 2024). Recent models further connect these representations to metric scene reconstruction and 3D interaction dynamics (Yang et al., 2026; Huang et al., 2026). Reconstruction in the robot frame uses visual observations and robot configuration to align geometry with the embodiment, as in Robo3R, which predicts metric point clouds from RGB images and robot states (Yang et al., 2026). Learning interaction dynamics instead requires a representation of both the scene and the applied action: PointWorld predicts future point displacements from calibrated RGB-D observations and robot actions represented as 3D point flows (Huang et al., 2026). Its formulation begins with an observed point cloud and forecasts the response to an action sequence, without prior visual frames or initial scene velocities as input. EVGT jointly reconstructs scene geometry and observed object motion from multi-view RGB sequences, estimating camera poses, depth, dense

tracks, and object states within a shared model. For this perception task, the formulation reduces dependence on auxiliary inputs: depth, camera calibration, robot states, and action commands are not required at inference. Relative to the point-flow output of PointWorld, EVGT also provides explicit object 6D pose, size, oriented boxes, masks, and visibility, making both dense geometry and compact object state directly available in a common metric frame.

Visual Geometry Foundation Models. Visual geometry foundation models learn transferable geometric priors that support feed-forward reconstruction of cameras and dense scene structure from uncalibrated images (Wang et al., 2024; 2025a;c). Recent developments broaden metric reconstruction across input configurations and extend inference to streaming observations through persistent spatial or temporal memory (Keetha et al., 2025; Wang et al., 2025b; Wu et al., 2025; Zhuo et al., 2025). Domain-specific models further adapt these representations to the observation structure of physical environments. In driving, the DVGT series models spatial and temporal relations in multi-camera video, supporting metric reconstruction in both batch and streaming settings (Zuo et al., 2025; 2026). These developments establish a foundation for scene geometry estimation across views and time. EVGT extends this formulation to embodied scenes by jointly modeling changing relative camera poses, dense scene geometry, and explicit object states, including 6D pose and motion.

Dynamic Geometry and 3D Motion. Dynamic reconstruction methods extend geometric estimation to videos containing both camera and scene motion, using temporal correspondences to recover evolving 3D structure (Zhang et al., 2025b; Sucar et al., 2025; 2026). Dense point maps and spatiotemporal trajectory representations further support reconstruction and tracking of moving surfaces across observations (Feng et al., 2025; Karhade et al., 2025; Zhang et al., 2025a; Luo et al., 2026). These representations describe motion at the level of individual points, while object pose estimation provides a compact description of rigid motion and object extent (Wen et al., 2024). EVGT connects these two levels through a shared representation of scene geometry, dense tracks, and temporal object features. The resulting outputs describe both the motion of visible surface points and the translation and rotation of the object as a whole, together with masks and visibility across views.

3 METHOD

3.1 OBJECT-CENTRIC 3D SCENE RECONSTRUCTION

General visual geometry foundation models estimate camera poses and dense scene geometry directly from uncalibrated images (Wang et al., 2024; 2025a;c). They encode images into tokens, model geometric relations with attention, and decode features into camera parameters, depth maps, and point maps. Given an image collection $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^N$, this process can be written as

$$\{\mathbf{C}_i, \mathbf{D}_i, \mathbf{P}_i\}_{i=1}^N = f_\theta(\mathcal{I}), \tag{1}$$

where $\mathbf{C}_i \in SE(3)$ denotes a camera pose, $\mathbf{D}_i \in \mathbb{R}^{H\times W}$ a depth map, and $\mathbf{P}_i \in \mathbb{R}^{H\times W\times 3}$ a point map in a shared reference frame. Each point-map entry assigns a continuous 3D position to an image pixel, providing a dense description of visible scene structure.

Embodied observations introduce additional structure across cameras, time, and objects. External and wrist cameras can move relative to one another, so each view requires a camera pose that may change over time. Objects can also move independently of the cameras, and their geometry, pose, and motion must remain associated across observations. We therefore formulate embodied perception as joint reconstruction of scene geometry and object state from a video clip

$$\mathcal{O}_{1:T} = \{\mathbf{I}_{t,v} \mid t = 1, \dots, T,\ v = 1, \dots, V\}, \tag{2}$$

with T time steps and V views, without calibration, depth, robot states, or actions as inputs.

We use the coordinate system of a designated camera at the first time step as a fixed reference frame, denoted by `ref0`. Camera poses $\mathbf{C}_{t,v}$ map local camera coordinates into this frame. Point maps, object poses, and 3D motion are expressed in the same frame, while depth is associated with each input camera. This formulation accommodates changing relative camera poses without requiring a fixed robot base frame. The scene representation is

$$\mathcal{G}_{1:T} = \{\mathbf{C}_{t,v}, \mathbf{D}_{t,v}, \mathbf{P}_{t,v}, \mathbf{U}_{t,v}\}_{t,v}, \tag{3}$$

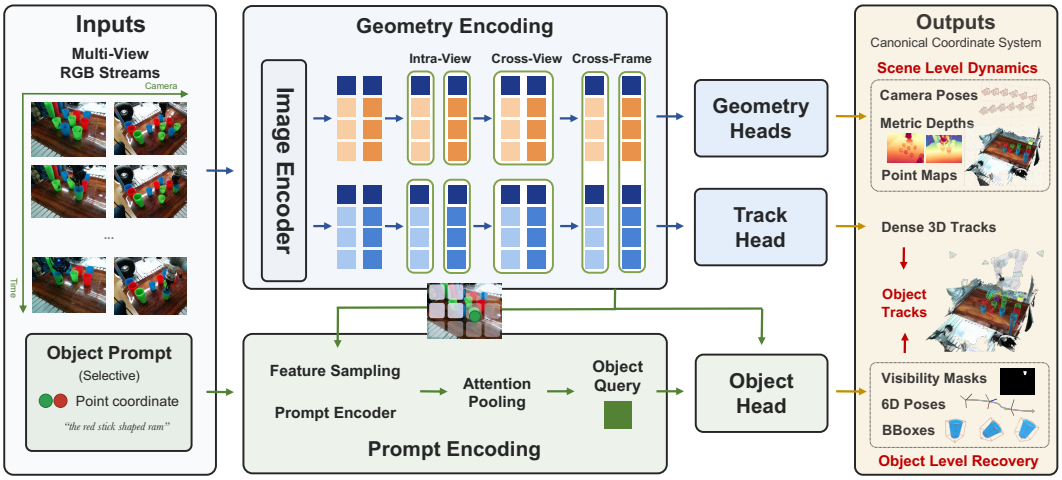


Figure 2: **EVGT architecture.** The image encoder and spatiotemporal geometry backbone extract features from multi-view RGB sequences. Geometry and motion heads estimate camera poses, metric depth, point maps, and dense 3D tracks. The object readout aggregates scene and motion features across views and time to decode 6D pose, size, oriented boxes, masks, and visibility. Optional prompts condition motion features and object queries. Training first learns scene geometry and motion, then adds object supervision while continuing to optimize the shared representation.

where $\mathbf{U}_{t,v}$ denotes dense 3D displacement from a selected query frame to time t, indexed by pixels in query view v. For an object or rigid part o, we additionally estimate

$$\mathcal{S}_t^o = \left(\mathbf{T}_t^o, \mathbf{s}^o, \mathbf{B}_t^o, \mathbf{M}_{t,1:V}^o, \mathbf{v}_{t,1:V}^o\right), \tag{4}$$

where $\mathbf{T}_t^o = [\mathbf{R}_t^o, \mathbf{t}_t^o] \in SE(3)$ is the 6D pose, $\mathbf{s}^o \in \mathbb{R}_+^3$ the size shared across the clip, and $\mathbf{B}_t^o \in \mathbb{R}^{8\times 3}$ the oriented box corners. The per-view mask $\mathbf{M}_{t,v}^o$ and visibility $\mathbf{v}_{t,v}^o$ describe its image support and observability. An optional prompt q specifies the target. The full task is

$$(\mathcal{G}_{1:T}, \mathcal{S}_{1:T}^o) = f_{\text{EVGT}}(\mathcal{O}_{1:T}; q), \tag{5}$$

with q omitted when no target specification is provided. Dense tracks describe surface motion, while the pose sequence provides the translation and rotation of the object as a whole.

3.2 EVGT: Embodied Visual Geometry Transformer

We implement this formulation with an image encoder, a spatiotemporal geometry transformer, a motion decoder, an object readout, and multiple prediction heads, as shown in Figure 2. The geometry pathway establishes a representation of the scene across views and time, and object queries aggregate scene and motion features into a temporal object representation. Optional prompt features provide target information through the motion pathway and the object queries.

Visual Encoding. DINOv3 ViT-L/16 (Siméoni et al., 2025) encodes each image into tokens:

$$\mathbf{F}_{t,v} = \mathcal{E}(\mathbf{I}_{t,v}). \tag{6}$$

Spatiotemporal Interaction. Following the factorized attention design of DVGT (Zuo et al., 2025), the geometry transformer uses 24 blocks that alternate attention within images, across synchronized views, and across frames. Writing these operations as $\mathcal{A}_{\text{img}}$, $\mathcal{A}_{\text{view}}$, and $\mathcal{A}_{\text{time}}$, each block is

$$\mathbf{F}^{\ell+1} = \mathcal{A}_{\text{time}}^{\ell}\left(\mathcal{A}_{\text{view}}^{\ell}\left(\mathcal{A}_{\text{img}}^{\ell}(\mathbf{F}^{\ell})\right)\right). \tag{7}$$

Attention within each image retains local geometric detail, attention across views aggregates complementary observations, and attention across frames models temporal correspondence. This organization lets features gather evidence as wrist viewpoints change and objects move or become occluded. A three-axis rotary position encoding indexes frame, patch row, and patch column.

Optional Prompt Encoding. When supplied, language or point prompts provide a target vector $\mathbf{g} \in \mathbb{R}^{256}$ for the object queries. For point prompts, visual features sampled at the annotated positions are combined with Fourier-encoded coordinates, relative time, camera slot, and click polarity,

and a learned seed attends to the resulting tokens. A frozen SAM3 text encoder (Carion et al., 2025) provides language features, which are projected to the same dimension and interact with scene tokens. The two query representations are averaged when both prompt modalities are present.

Motion Decoder. The motion decoder produces track features $\mathbf{F}^{\mathrm{trk}}$ for dense 3D displacement prediction. These features also feed motion information to the object readout. With a prompt, its target features condition the motion decoder through feature-wise affine adapters,

$$\widetilde{\mathbf{F}}_{\ell}^{\mathrm{trk}} = (1 + \boldsymbol{\gamma}_{\ell}(\mathbf{g})) \odot \mathbf{F}_{\ell}^{\mathrm{trk}} + \boldsymbol{\beta}_{\ell}(\mathbf{g}), \tag{8}$$

where the zero-initialized adapter outputs preserve the pretrained motion features.

Object Readout. We combine the final scene tokens and motion-decoder tokens into an object token bank. Layer normalization and linear projection map both sources to 256 dimensions, retaining position, time, view, and token-source encodings. Camera identity is retained as an observation slot in the object readout. At each frame, an object query attends to this bank across views through two cross-attention layers. A two-layer temporal transformer then exchanges information among the frame representations, producing object features $\mathbf{z}^{o}_{1:T}$. These features retain the evidence needed to estimate object state when visibility differs across views or time. A pooled clip representation summarizes attributes that remain constant across frames.

Geometry Heads. The camera head decodes the refined visual features as $\widehat{\mathbf{C}}_{t,v} = \mathcal{H}_{\mathrm{cam}}(\mathbf{F}^{L}_{t,v})$. Predicting a pose for each view and frame represents both motion of the camera system and changes in relative camera poses. The depth head predicts a metric depth map for each input image, $\widehat{\mathbf{D}}_{t,v} = \mathcal{H}_{\mathrm{depth}}(\mathbf{F}^{L}_{t,v})$. These pixel-aligned predictions describe scene depth in the corresponding camera frame. The point-map head predicts $\widehat{\mathbf{P}}_{t,v} = \mathcal{H}_{\mathrm{point}}(\mathbf{F}^{L}_{t,v})$, assigning each image pixel a metric 3D position in `ref0`. Point maps from different views and frames therefore describe visible geometry in a shared coordinate system. Additional confidence heads estimate the reliability of the dense depth and point-map predictions. The dense track head decodes the motion features into 3D displacements and confidence. For a query pixel $\mathbf{u}$ in view v at frame r, its trajectory is represented by

$$\widehat{\mathbf{X}}_{t,v}(\mathbf{u}) = \widehat{\mathbf{P}}_{r,v}(\mathbf{u}) + \widehat{\mathbf{U}}_{t,v}(\mathbf{u}). \tag{9}$$

Both terms use `ref0`, so trajectories describe scene motion in a fixed frame.

Object Heads. The pose head decodes $\mathbf{z}^{o}_{t}$ into translation and a continuous six-parameter rotation representation, which is converted into a rotation matrix. The size head predicts positive metric extents from the clip representation. The predicted size is shared across frames to represent the fixed extent of a rigid object. Together, pose and size determine the oriented box analytically:

$$\widehat{\mathbf{b}}^{o}_{t,k} = \widehat{\mathbf{R}}^{o}_{t}\left(\tfrac{1}{2}\widehat{\mathbf{s}}^{o} \odot \mathbf{c}_{k}\right) + \widehat{\mathbf{t}}^{o}_{t}, \qquad \mathbf{c}_{k} \in \{-1, 1\}^{3}, \tag{10}$$

where the eight sign combinations index the box corners. The visibility head combines temporal object features with context from each view to predict object visibility. The mask decoder matches an object-conditioned query to dense visual features augmented with RGB detail, producing a mask for each view and frame. An additional head produces an L2-normalized 128-dimensional instance embedding from the pooled clip representation. Its contrastive objective is disabled in the evaluated configuration. To obtain object-surface tracks, we sample query pixels from the predicted mask, keep valid positive point prompts, and gather the trajectories in Eq. (9). The resulting surface tracks complement the object-center trajectory $\{\widehat{\mathbf{t}}^{o}_{t}\}^{T}_{t=1}$.

3.3 Training

Training Data. Training combines simulated robot interactions from LIBERO, ManiSkill3, MimicGen, and RLBench (Liu et al., 2023; Mandlekar et al., 2023; James et al., 2020), object-pose datasets such as DexYCB and HO3D (Chao et al., 2021; Hampali et al., 2020), real robot videos with mask supervision from RoboCOIN, RH20T, FMB, and AgiBot (Fang et al., 2024), and geometry replay sources such as PointOdyssey (Zheng et al., 2023). Table 7 lists all sources and quotas.

Data Annotation. We use native ground truth where available and construct the remaining targets through source-specific pipelines. For real videos, calibrated RGB-D, MoGe-2 depth (Wang et al., 2025b), SAM masks (Carion et al., 2025), and 4RC temporal geometry (Luo et al., 2026) yield depth,

point maps, and quality-gated tracks. Samples without a full object pose receive weak translation targets from back-projected mask depth. All targets are expressed in `ref0`, each loss uses only valid targets, and Appendix B details each annotation pipeline.

Training Objectives. Stage 1 learns scene geometry and dense motion, and stage 2 adds object supervision while jointly optimizing the shared representation. The combined objective is

$$\begin{aligned}\mathcal{L} = \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{trk}}\mathcal{L}_{\text{trk}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{size}}\mathcal{L}_{\text{size}} \\ + \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{vis}}\mathcal{L}_{\text{vis}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}},\end{aligned} \tag{11}$$

where the weights λ balance the terms and each term is evaluated only where its target is valid. Geometry supervision covers camera poses, depth, and point maps, and track supervision covers 3D displacement and confidence. Object pose supervision combines Smooth-L1 translation regression, symmetry-aware rotation supervision, and relative rotation across frames. Size supervision penalizes log-space errors in metric extents, box supervision measures bidirectional corner distances, and the mask objective combines binary cross-entropy and Dice losses. The auxiliary term adds foreground-track, point-prompt, and surface-in-box constraints, and Appendix C defines every term.

Training Stages. Stage 1 runs for 600,000 steps on sources with native or carefully annotated geometry. Stage 2 continues for 28,250 updates, sampling text-only, point-only, and combined prompts with probabilities 0.4, 0.4, and 0.2. The DINO encoder and camera head are frozen, while the last four blocks of each attention type, the motion decoder, and the prediction heads are optimized with module-specific learning rates. Further details are given in Appendix A.

4 Experiments

For scene geometry and motion, we evaluate EVGT on six robot-operation sources using the same frozen RGB inputs as VGGT, 4RC, and Pi3. The scene tables report point-map reconstruction, depth, camera pose, and dynamic 3D point tracking by source, and a separate object table reports center and rotation diagnostics. All four scene tables use the same EVGT weights. The downstream manipulation experiment reports closed-loop policy success on LIBERO-Plus and RoboCasa24.

4.1 Implementation Details

For the scene-geometry tables, all methods receive frozen 224×224 RGB inputs, with each baseline retaining its native internal preprocessing. The four core robot-operation sources use single-camera and synchronized two-camera clips sampled at matching time points. Where duration and timestamps permit, clips contain 24 distinct time points at 4 Hz, and shorter clips retain their native timing. RH20T and FMB use their calibrated real-robot sequences at native timing.

4.2 Benchmark and Metrics

Evaluation Sources. LIBERO, ManiSkill3, MimicGen, and RLBench contribute 50 candidate episodes each, and RH20T and FMB contribute another 50 each. The four core sources and FMB average single- and two-camera results within an episode, while RH20T uses its single-camera protocol. FMB includes 32 episodes present in the training and development indices, so its results measure in-distribution retention rather than unseen-source transfer. The object-state table evaluates 50 windows each from the four core sources, DexYCB, and HO3D (Appendix D).

Baselines. VGGT (Wang et al., 2025a), 4RC (Luo et al., 2026), and Pi3 (Wang et al., 2025c) are run on the same frozen RGB inputs as EVGT. For object state, we compare with SAM 2.1 followed by SAM 3D Objects and with FoundationPose (Wen et al., 2024). Both use SAM 2.1 target masks, and FoundationPose also uses depth, camera intrinsics, and an object mesh, whereas EVGT uses RGB and a first-frame click.

Prompt Conditions. EVGT accepts text descriptions, image clicks, and their combination. The object table uses first-frame clicks, while text-only and text-plus-click prompts are left for later evaluation. Relative object motion and full-pose metrics remain outside this table.

Metrics. Point maps use Accuracy and Completeness after one sequence-level Sim(3) alignment, and depth uses AbsRel and δ_1 with one scale correction per sequence. The core sources use ATE

Table 1: **Point-map reconstruction by source.** Lower is better (m). Best in bold.

	LIBERO		ManiSkill3		MimicGen		RLBench		RH20T		FMB	
Method	Acc.	Comp.	Acc.	Comp.	Acc.	Comp.	Acc.	Comp.	Acc.	Comp.	Acc.	Comp.
VGGT	0.114	0.141	0.349	0.286	0.114	0.136	0.238	0.200	0.049	0.071	0.0268	0.0289
4RC	0.107	0.143	0.317	0.270	0.106	0.166	0.247	0.238	0.048	0.064	0.0156	0.0166
Pi3	0.119	0.140	0.309	0.208	0.105	0.109	0.221	0.180	0.045	0.058	0.0283	0.0342
EVGT	**0.036**	**0.037**	**0.125**	**0.124**	**0.084**	**0.099**	**0.102**	**0.143**	**0.036**	**0.048**	**0.0078**	**0.0082**

Table 2: **Depth prediction by source.** Lower AbsRel and higher δ_1 are better. Best in bold.

	LIBERO		ManiSkill3		MimicGen		RLBench		RH20T		FMB	
Method	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
VGGT	0.463	0.531	0.459	0.440	1.303	0.299	0.553	0.491	0.085	0.929	0.069	0.961
4RC	0.354	0.563	0.335	0.519	1.255	0.301	0.509	0.503	0.110	0.897	0.052	0.975
Pi3	0.297	0.624	0.261	0.638	1.028	0.356	0.383	0.579	0.069	0.949	0.072	0.949
EVGT	**0.162**	**0.833**	**0.220**	**0.757**	**0.296**	**0.688**	**0.191**	**0.798**	**0.032**	**0.982**	**0.016**	**0.998**

Table 3: **Camera-pose estimation by source.** ATE is in meters, Drift-t in centimeters, and RPE-r and Drift-r in degrees. Lower is better. Best in bold.

	LIBERO		ManiSkill3		MimicGen		RLBench		RH20T		FMB	
Method	ATE	RPE-r	ATE	RPE-r	ATE	RPE-r	ATE	RPE-r	Drift-t	Drift-r	Drift-t	Drift-r
VGGT	0.069	3.984	0.068	4.179	0.109	7.811	0.078	4.984	0.124	0.120	0.938	0.692
4RC	0.087	4.712	0.080	2.084	0.150	**3.741**	0.086	3.178	**0.031**	0.048	0.223	0.190
Pi3	0.051	**3.369**	0.043	**1.948**	0.104	3.840	**0.044**	**2.318**	0.248	0.203	0.432	0.299
EVGT	**0.048**	3.373	**0.038**	3.730	**0.059**	4.652	0.049	5.986	0.060	**0.028**	**0.028**	**0.020**

and adjacent-frame RPE-r, and the fixed RH20T and FMB cameras use translation and rotation drift from the first frame. In the two-camera protocol, the moving wrist and fixed external cameras jointly determine the alignment and scores, so these values summarize multi-camera estimation rather than wrist-camera motion alone. Dynamic tracks use protocol-B EPE and APD over queries that move at least 2 cm, and object state uses center and absolute rotation errors. Scores are averaged within each episode and reported by source, and Appendix E gives the full definitions.

4.3 Main Results

Point-Map Reconstruction. Table 1 reports Accuracy and Completeness over 50, 43, 50, 50, 50, and 50 episodes in source order. EVGT has the lowest Accuracy and Completeness on all six sources. We next evaluate the depth predictions that accompany these point maps.

Depth Prediction. Table 2 reports scale-corrected AbsRel and δ_1 on the same episodes. EVGT obtains the lowest AbsRel and highest δ_1 on all six sources. Its raw AbsRel is also the lowest on RH20T and FMB, at 0.034 and 0.017 compared with at least 0.443 and 1.180 for the baselines. We next examine whether the representation maintains camera geometry across time and viewpoints.

Camera-Pose Estimation. Table 3 separates multi-camera ATE and RPE-r on the four core sources from fixed-camera drift on RH20T and FMB, over 40, 39, 41, 31, 50, and 50 episodes in source order. EVGT has the lowest ATE on LIBERO, ManiSkill3, and MimicGen, while Pi3 leads RLBench and adjacent-frame rotation results vary by source. For fixed cameras, EVGT has the lowest rotation drift on RH20T and FMB and the lowest translation drift on FMB, while 4RC leads RH20T translation drift. We next assess whether the scene representation also preserves dynamic 3D point motion.

Dynamic 3D Point Tracking. Table 4 reports protocol-B EPE and APD on the four core sources over common episodes for EVGT, VGGT, and 4RC. Common N counts valid and candidate episodes under both camera protocols, each episode averages its single- and two-camera scores, and the final row averages the four source means equally. Pi3 has no dynamic 3D-track output. EVGT has the lowest EPE on all four sources and the highest APD on LIBERO, ManiSkill3, and MimicGen, while 4RC has the highest APD on RLBench.

Table 4: **Dynamic 3D point tracking.** Cells show protocol-B EPE (m) ↓ / APD ↑. Best in bold.

Source	Common N	EVGT	VGGT	4RC
LIBERO	50/50	**0.0616 / 0.3948**	0.1199 / 0.1490	0.1312 / 0.1061
ManiSkill3	43/43	**0.0939 / 0.2632**	0.1291 / 0.1743	0.1085 / 0.2326
MimicGen	45/45	**0.0815 / 0.2683**	0.1163 / 0.1372	0.1166 / 0.1489
RLBench	44/45	**0.1043** / 0.2045	0.1392 / 0.1347	0.1059 / **0.2412**
Four-source mean	182/183	**0.0853 / 0.2827**	0.1261 / 0.1488	0.1156 / 0.1822

Table 5: **Object-center error (Ctr., cm) and absolute rotation error (Rot., degrees).** Lower is better. Best in bold. N/A marks no rotation output in the canonical object frame.

	LIBERO		ManiSkill3		MimicGen	
Method	Ctr. ↓	Rot. ↓	Ctr. ↓	Rot. ↓	Ctr. ↓	Rot. ↓
SAM2+SAM3D	166.31	N/A	128.03	N/A	173.74	N/A
FoundationPose	**21.52**	130.12	**12.17**	103.46	**10.99**	99.84
EVGT	26.06	**55.16**	19.42	**38.57**	23.08	**82.51**
	RLBench		DexYCB		HO3D	
Method	Ctr. ↓	Rot. ↓	Ctr. ↓	Rot. ↓	Ctr. ↓	Rot. ↓
SAM2+SAM3D	122.62	N/A	20.76	N/A	56.32	N/A
FoundationPose	**6.93**	109.34	**10.06**	**61.07**	6.40	**45.72**
EVGT	45.03	**98.05**	17.51	123.55	**6.33**	104.75

Object-State Evaluation. Table 5 reports object-center and rotation errors across six sources. SAM2+SAM3D feeds SAM 2.1 masks to SAM 3D Objects, and FoundationPose uses proxy meshes for the four simulation sources and CAD models for DexYCB and HO3D. The EVGT values use the latest native-input, `ref0` scoring run, while SAM2+SAM3D and the DexYCB and HO3D FoundationPose scores come from an earlier 224×224 diagnostic. FoundationPose returns poses for 50/50, 50/50, 45/50, and 48/50 simulation windows, and its means use successful cases. FoundationPose, which also uses depth and object meshes, has the lowest center error on five sources, while EVGT has the lowest rotation error on the four simulation sources. Since inputs and scoring differ across methods, these are conditional diagnostics rather than a matched-input ranking.

4.4 Downstream Manipulation Transfer

Following the frozen geometric-feature transfer setting of VGGT-Ω (Wang et al., 2026), we test whether EVGT features help a vision-language-action policy on LIBERO-Plus and RoboCasa24. We augment $\pi_{0.5}$ (Physical Intelligence et al., 2025) with scene patches and register tokens from a frozen EVGT encoder. The original $\pi_{0.5}$ visual tokens query these features through a zero-initialized cross-attention residual. The EVGT encoder and the original $\pi_{0.5}$ visual tower remain frozen, while the language backbone, action expert, and fusion adapter are fine-tuned. LIBERO-Plus and RoboCasa24 use two and three camera views, each with its own $\pi_{0.5}$ initialization and training data.

Table 6 reports closed-loop task success. On LIBERO-Plus, the $\pi_{0.5}$ starting checkpoint succeeds in 8,492 of 10,030 rollouts (84.67%), while $\pi_{0.5}$+EVGT after 24k adaptation steps succeeds in 8,657 (86.31%), a gain of 1.65 points. On its 1,599 camera-viewpoint-change rollouts, success is 77.67% and 90.37%, respectively. Across the four LIBERO suites, Spatial is unchanged, Object decreases by 1.75 points, and Goal and Long improve by 4.67 and 3.49 points. On RoboCasa24, $\pi_{0.5}$+EVGT after 15k steps succeeds in 710 of 1,200 rollouts over 24 tasks (59.17%), compared with 55.1% for the fine-tuned RGB-only $\pi_{0.5}$ baseline of SeeR-VLA (Lee et al., 2026).

4.5 Qualitative Results

Figure 3 shows that EVGT reconstructs spatially consistent scene geometry across camera viewpoints, with the recovered surfaces following the observed workspace layout. In the selected simulated and real hand–object sequences of Fig. 4, its 3D boxes and pose axes follow the reference as the target translates and rotates, showing accurate and temporally consistent object states.

Table 6: **Closed-loop manipulation success rate (SR, %).** Higher is better. Best in bold.

Benchmark	Policy	Overall SR ↑	Camera Viewpoints SR ↑
LIBERO-Plus	$\pi_{0.5}$ starting checkpoint	84.67	77.67
	$\pi_{0.5}$ + EVGT	**86.31**	**90.37**
RoboCasa24	RGB $\pi_{0.5}$ (fine-tuned)	55.12	-
	$\pi_{0.5}$ + EVGT	**59.17**	-

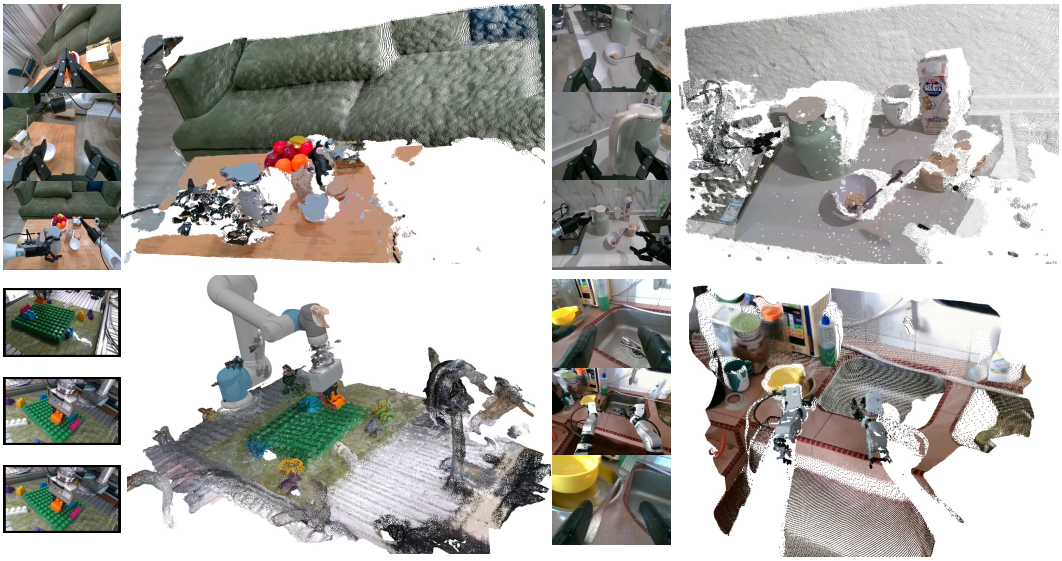


Figure 3: **Multi-view scene-geometry examples on core robot-video episodes.**

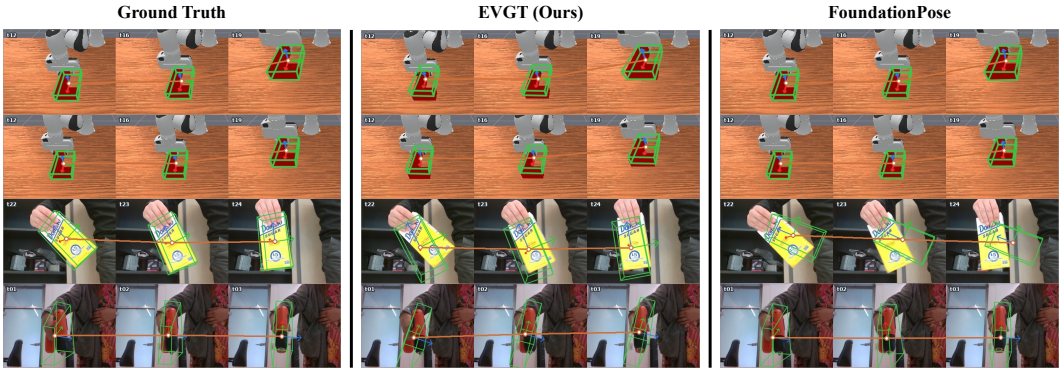


Figure 4: **Object-state examples.** Left: ground truth. Middle: EVGT. Right: FoundationPose.

5 Conclusion

We have presented EVGT, an Embodied Visual Geometry Transformer that serves as a metric-scale geometry foundation model for robotic scenes. EVGT integrates multi-view RGB observations with temporally varying relative camera poses to jointly reconstruct scene geometry, object 6D pose, and object motion in a common metric reference frame. The model combines visual tokens and object queries through spatiotemporal interaction and is trained in two stages that pair heterogeneous scene-geometry supervision with partially labeled object data. The evaluation covers scene reconstruction, camera trajectories, dense motion, object-state diagnostics, and downstream manipulation success.

Limitations. EVGT treats each target as a rigid object with a fixed extent, so articulated and deformable objects are not modeled, and object identity is not maintained across clips. Its absolute rotation error remains higher than that of FoundationPose on DexYCB and HO3D, where FoundationPose also uses depth and CAD models. The evaluation also relies on first-frame clicks and a limited set of real-robot sources, and part of FMB overlaps with the training indices.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and manuscript organization. All technical claims, equations, citations, experiments, and reported results are verified by the authors.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Sections 3.2, 3.3 and 4.1 describe the model, training, and implementation, and Appendices A, C and E detail the training configuration, loss terms, and metrics. We will release the complete source code and model weights.

REFERENCES

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. URL https://arxiv.org/abs/2307.15818.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, et al. SAM 3: Segment anything with concepts. *arXiv preprint*, 2025.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stanley Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ruijie Fang et al. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In *Robotics: Science and Systems (RSS)*, 2024.

Haiwen Feng, Junyi Zhang, Angjoo Kanazawa, et al. St4RTrack: Simultaneous 4d reconstruction and tracking in the world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. *arXiv preprint arXiv:2306.17817*, 2023.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. *arXiv preprint arXiv:2306.14896*, 2023.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HO-3D: A large scale dataset for joint 3d hand and object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. *Autonomous Robots*, 2013. doi: 10.1007/s10514-012-9321-0. URL https://octomap.github.io.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Fei-Fei Li. PointWorld: Scaling 3D world models for in-the-wild robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Huang_PointWorld_Scaling_3D_World_Models_for_In-The-Wild_Robotic_Manipulation_CVPR_2026_paper.pdf.

Stephen James, Zicong Ma, David Rojas Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.

Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4D: Unified feed-forward metric 4d reconstruction. *arXiv preprint arXiv:2512.10935*, 2025.

Nikhil Varma Keetha, Norman Müller, Johannes L. Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, et al. MapAnything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Byungkun Lee, Dongyoon Hwang, Dongjin Kim, Hojoon Lee, Hyunseung Kim, Jaegul Choo, and Minho Park. See like a robot: Robot-centric pointmaps for VLA models. *arXiv preprint arXiv:2607.11498*, 2026. URL https://arxiv.org/abs/2607.11498.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

Yihang Luo et al. 4RC: 4d reconstruction via conditional querying anytime. In *International Conference on Machine Learning (ICML)*, 2026.

Kevin M. Lynch and Frank C. Park. *Modern Robotics: Mechanics, Planning, and Control*. Cambridge University Press, 2017. doi: 10.1017/9781316661239.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. MimicGen: A data generation system for scalable robot learning using human demonstrations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

Physical Intelligence, Kevin Black, Noah Brown, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic point maps: A versatile representation for dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: 4d video reconstruction with dynamic point maps. *arXiv preprint arXiv:2601.09499*, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 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 (CVPR)*, 2026. URL https://arxiv.org/abs/2605.15195.

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. *arXiv preprint arXiv:2507.02546*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13328*, 2025c.

Bowen Wen, Wei Yang, Jan Kautz, and Stanley Birchfield. FoundationPose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025. URL https://arxiv.org/abs/2507.02863.

Sizhe Yang, Linning Xu, Hao Li, Juncheng Mu, Jia Zeng, Dahua Lin, Jiangmiao Pang, et al. Robo3R: Enhancing robotic manipulation with accurate feed-forward 3d reconstruction. *arXiv preprint arXiv:2602.10101*, 2026.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle Barral, Raia Hadsell, et al. Efficiently reconstructing dynamic scenes one D4RT at a time. *arXiv preprint arXiv:2512.08924*, 2025a.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, and Deqing Sun. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations (ICLR)*, 2025b.

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 (ICCV)*, 2023.

Yifeng Zhu, Zhenyu Jiang, Peter Stone, and Yuke Zhu. GROOT: Learning generalizable manipulation policies with object-centric 3d representations. In *Conference on Robot Learning (CoRL)*, 2023.

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

Sicheng Zuo, Zixun Xie, Wenzhao Zheng, Shaoqing Xu, Fang Li, Shengyin Jiang, Long Chen, Zhi-Xin Yang, and Jiwen Lu. DVGT: Driving visual geometry transformer. *arXiv preprint arXiv:2512.16919*, 2025.

Sicheng Zuo, Zixun Xie, Wenzhao Zheng, Shaoqing Xu, Fang Li, Hanbing Li, Long Chen, Zhi-Xin Yang, and Jiwen Lu. DVGT-2: Vision-geometry-action model for autonomous driving at scale. *arXiv preprint arXiv:2604.00813*, 2026.

Table 7: **Stage-2 data mixture layout.**

Group	Sources	Primary supervision
Simulation	LIBERO, ManiSkill3, MimicGen, RLBench	camera, depth, point map, tracks, mask, pose, box/size
Real 6D	DexYCB, HO3D, ACE-Data-0, ADT, TACO, HouseCat6D, MOVi-E, PACE-Sim, SOPE	object 6D pose, true size
Real Mask	RoboCOIN (tracked/untracked), RH20T, FMB, AgiBot, ProxyPose position annotations	masks, mask-depth weak translation, and ProxyPose position tracks (weight 0.25, AgiBot at 4 Hz and 512×288)
Replay	RoboCOIN, RoboMind2, HyperSim, PointOdyssey, Dynamic Replica	inherited stage-1 supervision
Humanoid	G1-ReplicaCAD, HIW500, Humanoid Everyday	held-out generalization probes

Appendix

A Training Configuration

Stage 1 trains for 600,000 optimization steps on sources with native ground truth and carefully annotated real videos. The prompted stage-2 continuation starts from the resulting geometry checkpoint and runs for 28,250 updates. The DINO encoder and camera head are frozen. The last four blocks of each aggregator type (intra-view, cross-view, cross-frame), motion decoder, track head, zero-initialized prompt adapters, object head, depth head, and point head are trained with module-specific learning rates. Prompts are sampled as 40% point-only, 40% text-only, and 20% combined. Clicks are sampled from ground-truth masks, and text descriptions are paired with instances. A scale-calibration branch uses window statistics on raw depth, a relative-unit head, and a zero-initialized scale correction. It is implemented but excluded from all reported results.

Table 7 lists the stage-2 data mixture. Within each block of 64 drawn samples, simulation data supplies the fullest set of labels, real 6D-pose data supplies object pose and size, real mask data supplies masks, mask-depth weak translation targets, and ProxyPose position annotations, and replay data refreshes stage-1 supervision. Humanoid sources are outside this quota and reserved as generalization probes. Mask-depth weak translation targets supervise only query-frame `ref0` translation obtained by back-projecting target-mask depth. Rotation, size, and box are invalid for those samples. ProxyPose separately provides precise multi-frame object-position labels and associated masks at training weight 0.25. Dataset splits follow the stage-1 partition. DexYCB and HO3D are held out by session and rig, while 2.5% of AgiBot and ProxyPose is held out by episode and case hash.

B Training Data and Annotation

Stage 1 uses datasets with native geometric ground truth and carefully annotated real-robot videos, including RoboCOIN and AgiBot. The stage-2 object-pose group includes DexYCB, HO3D, ACE-Data-0, ADT, TACO, HouseCat6D, MOVi-E, PACE-Sim, and SOPE, with pose and size labels. Real mask supervision comes from RoboCOIN, RH20T, FMB, and AgiBot, with ProxyPose position labels on some videos. Replay samples from RoboCOIN, RoboMind2, HyperSim, PointOdyssey, and Dynamic Replica retain scene-geometry supervision. In each 64-sample stage-2 block, the simulation, object-pose, real-mask, and replay groups contribute 20, 16, 16, and 12 samples.

The sources in Table 7 provide different levels of supervision. We use native ground truth where available and construct additional targets through source-specific processing. Simulation sources contribute scene geometry, tracks, object masks, pose, and size, while pose-annotated sources contribute their available object-state ground truth. Camera and object poses, point maps, and 3D tracks are expressed in the `ref0` frame of the first reference camera. Geometric validity, target-object masks, and track validity are recorded separately.

Real videos follow source-specific annotation pipelines. RH20T and FMB use native RGB-D observations, with camera-calibration checks and SAM-based robot masks defining valid geometry pixels (Carion et al., 2025). For fixed-camera RoboCOIN videos, MoGe-2 (Wang et al., 2025b) estimates depth, and constraints from static background regions stabilize the geometry over time. Target masks and quality-screened correspondences yield camera-wise tracks. AgiBot synchronizes three cameras by timestamp and calibrates MoGe-2 predictions with native head-camera depth. SAM masks identify robot pixels for geometric filtering and held objects for depth repair. Cross-view calibration and 4RC (Luo et al., 2026) provide temporal geometric evidence for the delivered dense depth and point maps, and tracks pass a separate quality gate.

For samples with a target mask and valid depth but no full object pose, we back-project the query-frame target region to form a weak translation target. ProxyPose provides precise multi-frame object-position labels with associated masks at training weight 0.25. Replay samples retain their available geometry and track labels, while humanoid sources serve as separate generalization probes. Each loss is computed only over the frames, views, and pixels for which its target is valid.

C Loss Definitions

We expand the terms in Eq. (11) below. A hat denotes a prediction and an asterisk denotes its supervision target. For each term k, $\mathcal{V}_k$ contains only elements with valid annotations, and $\mathcal{R}_{\mathcal{V}_k}$ denotes reduction over those elements, including any source-specific supervision weights. Missing annotations contribute no loss. Internal weights α differ from the task weights λ in Eq. (11).

Scene Geometry. The geometry objective combines camera, depth, and point-map supervision:

$$\begin{aligned}\mathcal{L}_{\text{geo}} = \alpha_{\text{cam}}\mathcal{R}_{\mathcal{V}_{\text{cam}}}[\ell_{\text{cam}}(\widehat{\mathbf{C}}, \mathbf{C}^*)] \\ + \alpha_{\text{depth}}\mathcal{R}_{\mathcal{V}_{\text{depth}}}[\ell_{\text{depth}}(\widehat{\mathbf{D}}, \mathbf{D}^*; \widehat{c}^D)] \\ + \alpha_{\text{point}}\mathcal{R}_{\mathcal{V}_{\text{point}}}[\ell_{\text{point}}(\widehat{\mathbf{P}}, \mathbf{P}^*; \widehat{c}^P)].\end{aligned} \tag{12}$$

Camera supervision compares the predicted and reference camera poses, depth supervision compares corresponding image pixels, and point-map supervision compares 3D positions in `ref0`. The symbols $\widehat{c}^D$ and $\widehat{c}^P$ denote the corresponding predicted confidence values. The residual operators retain the confidence treatment of the geometry objective.

Dense Tracks. Track supervision compares displacements for valid query–frame pairs:

$$\mathcal{L}_{\text{trk}} = \mathcal{R}_{\mathcal{V}_{\text{trk}}}[\ell_{\text{trk}}(\widehat{\mathbf{U}}_{t,v}(\mathbf{u}), \mathbf{U}^*_{t,v}(\mathbf{u}); \widehat{c}^{\text{trk}}_{t,v}(\mathbf{u}))]. \tag{13}$$

The reference displacement is measured from the query-frame 3D point, using the same reference frame as Eq. (9). The track residual includes supervision of displacement and confidence.

Object Pose. Pose supervision combines translation, symmetry-aware rotation, and relative rotation:

$$\mathcal{L}_{\text{pose}} = \alpha_t\mathcal{L}_t + \alpha_R\mathcal{L}_R + \alpha_{\Delta R}\mathcal{L}_{\Delta R}. \tag{14}$$

Translation applies Smooth-L1 to the residual $\mathbf{e}_t = \widehat{\mathbf{t}}^o_t - \mathbf{t}^{o*}_t$. For scalar e and $\beta > 0$, it is

$$\rho_\beta(e) = \begin{cases} e^2/(2\beta), & |e| < \beta, \\ |e| - \beta/2, & \text{otherwise}, \end{cases} \qquad \mathcal{L}_t = \mathcal{R}_{\mathcal{V}_t}[\rho_\beta(\mathbf{e}_t)], \tag{15}$$

where ρ_β acts elementwise on a vector. Let $\mathcal{S}^o$ denote the valid rotational symmetries of the target object. The symmetry-aware term compares the prediction with equivalent reference orientations:

$$\mathcal{L}_R = \mathcal{R}_{\mathcal{V}_R}\left[\min_{\mathbf{S}\in\mathcal{S}^o} d_R(\widehat{\mathbf{R}}^o_t, \mathbf{R}^{o*}_t\mathbf{S})\right]. \tag{16}$$

Here, d_R denotes the rotation residual, and the identity is the only symmetry for an asymmetric object. Relative-rotation supervision compares changes in orientation over valid frame pairs:

$$\mathcal{L}_{\Delta R} = \mathcal{R}_{\mathcal{V}_{\Delta R}}[d_{\Delta R}(\Delta\widehat{\mathbf{R}}^o_{r,t}, \Delta\mathbf{R}^{o*}_{r,t})]. \tag{17}$$

Mask-depth weak translation targets and ProxyPose labels enter only $\mathcal{L}_t$, not the rotation terms.

Object Size and Box. The size objective compares positive metric extents in log space:

$$\mathcal{L}_{\text{size}} = \mathcal{R}_{\mathcal{V}_{\text{size}}}[\ell_s(\log \widehat{\mathbf{s}}^o - \log \mathbf{s}^{o*})]. \tag{18}$$

Logarithms act componentwise, so the residual measures relative discrepancies. Let $\widehat{\mathcal{B}}_t^o$ and $\mathcal{B}_t^{o*}$ be the predicted and reference corner sets. The box loss compares them in both directions:

$$\mathcal{L}_{\text{box}} = \mathcal{R}_{\mathcal{V}_{\text{box}}}[d_{\rightarrow}(\widehat{\mathcal{B}}_t^o, \mathcal{B}_t^{o*}) + d_{\rightarrow}(\mathcal{B}_t^{o*}, \widehat{\mathcal{B}}_t^o)]. \tag{19}$$

Here, $d_{\rightarrow}$ denotes the directed corner-distance term. Predicted corners are computed from pose and size using Eq. (10), coupling box supervision to both predictions.

Masks and Visibility. The mask objective combines binary cross-entropy and Dice terms,

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}}. \tag{20}$$

For a foreground probability p and binary target y, the cross-entropy is

$$\ell_{\text{BCE}}(p, y) = -y \log p - (1 - y) \log(1 - p), \qquad \mathcal{L}_{\text{BCE}} = \mathcal{R}_{\mathcal{V}_{\text{mask}}}[\ell_{\text{BCE}}(\widehat{M}, M^*)]. \tag{21}$$

The Dice term penalizes disagreement between the predicted foreground region and the reference mask at the region level. Visibility supervision compares $\widehat{v}_{t,v}^o$ with its label:

$$\mathcal{L}_{\text{vis}} = \mathcal{R}_{\mathcal{V}_{\text{vis}}}[\ell_{\text{vis}}(\widehat{v}_{t,v}^o, v_{t,v}^{o*})]. \tag{22}$$

Mask and visibility validity are handled separately, so either annotation can contribute alone.

Auxiliary Supervision. The auxiliary loss sums foreground-track, click, and in-box terms:

$$\mathcal{L}_{\text{aux}} = \alpha_{\text{fg}}\mathcal{L}_{\text{fg}} + \alpha_{\text{click}}\mathcal{L}_{\text{click}} + \alpha_{\text{in}}\mathcal{L}_{\text{in}}. \tag{23}$$

Foreground-track supervision applies the track objective to valid points belonging to the target object, emphasizing motion of the selected foreground. Point-prompt consistency constrains the target prediction at annotated positive and negative clicks according to their polarity. The surface-in-box constraint penalizes mask-associated 3D surface points that lie outside the predicted oriented box. For a surface point $\mathbf{x}$, its object-frame coordinates and componentwise box violation are

$$\mathbf{y} = (\widehat{\mathbf{R}}_t^o)^\top(\mathbf{x} - \widehat{\mathbf{t}}_t^o), \qquad \mathbf{e}_{\text{in}} = \left[|\mathbf{y}| - \tfrac{1}{2}\widehat{\mathbf{s}}^o\right]_+, \tag{24}$$

where absolute value and $[\cdot]_+ = \max(\cdot, 0)$ act componentwise. The one-way constraint reduces a penalty on $\mathbf{e}_{\text{in}}$ over valid mask-associated surface points and does not require every part of the box to contain an observed point. The instance-embedding objective and pose–track consistency objective are disabled in the evaluated configuration and contribute no training loss.

D Benchmark Label Coverage

The main-table candidate set comprises 50 episodes each from LIBERO, ManiSkill3, MimicGen, RLBench, RH20T, and FMB. The first four are the core robot-operation sources, while RH20T and FMB supply calibrated fixed-camera real-robot sequences. Point and depth metrics use 50, 43, 50, 50, 50, and 50 episodes in source order (293 total), and the camera table uses 40, 39, 41, and 31 core-source episodes plus 50 each from RH20T and FMB (251 total). Protocol-B dynamic tracking has 50, 43, 45, and 45 paired candidate episodes on the core sources, of which 50, 43, 45, and 44 produce scores (182 total). The object table includes 50 scored windows each from LIBERO, ManiSkill3, MimicGen, RLBench, DexYCB, and HO3D. The HO3D windows span 24 sequences, including cases in the training index, and serve as diagnostics rather than unseen-episode transfer.

E Metric Definitions

Point Accuracy averages distances from predicted points to the reference surface, while Completeness measures the reverse direction, after one sequence-level Sim(3) alignment. Depth AbsRel is mean absolute relative error and δ_1 is the fraction of valid pixels within a depth ratio of 1.25. Both use one scale correction per sequence. For the four core sources, ATE measures camera-center error after one Sim(3) fit to valid cameras in a sequence, and RPE-r measures adjacent-frame rotation

error across valid camera streams. In the two-camera protocol, the moving wrist and fixed external cameras jointly determine the alignment and scores, so these values summarize multi-camera estimation rather than wrist-camera motion alone. Raw AbsRel on RH20T and FMB additionally exposes metric-scale behavior. For the fixed-camera RH20T and FMB sequences, Drift-t and Drift-r measure translation and rotation relative to the first frame, with the anchor frame omitted and translation scaled from native sensor depth. Protocol-B Dynamic Track EPE is the aligned 3D end-point error for queries with at least 2 cm of ground-truth world displacement. Its sequence-level Sim(3) is fitted from a fixed set of predicted/reference track correspondences. APD averages the fractions of those queries within 1, 2, 5, and 10 cm. Results are aggregated within each source, and the track table also reports the equal-weight mean of its four source means. Object-center error is the 3D distance between predicted and reference centers in centimeters, and absolute rotation error is reported in degrees. The object table mixes two input-grid and scoring protocols (Section 4.3).

F Prompt Conditions and Diagnostic Scope

The model accepts text descriptions, image clicks, and their combination. The current object table uses first-frame clicks and reports center and rotation diagnostics, while text-only and text-plus-click prompts form separate EVGT conditions for later evaluation. Relative object motion and harmonized CAD-axis and symmetry-based pose metrics remain outside the current table.