

ObjectWorld: Scaling Object-Centric World Models with A Large-Scale 3D Full-Object Flow Dataset

Haonan Chen^{1,2,3}, Sicheng Isabella Lu^{1,2}, Haitian Li^{1,2}, Jiajin Wu², Hexiang Wei², Jiawei Zhang^{1,2}, Xin Fei^{1,2}, Yiwen Hou^{1,2}, Chenrui Tie^{1,2}, Chongkai Gao^{1,2}, Sheng Wang², Lin Shao^{†1,2,3}



Fig. 1: **From diverse videos to executable full-object flow.** We construct a 1,210.1-hour dataset of 3D full-object flow from diverse real-world and simulated videos. An object-centric world model predicts dense motion over the complete object surface from image-and-language inputs, and the predicted flow guides downstream robot execution. The radial chart summarizes its composition across seven simulation datasets and three large-scale real-world video sources.

Abstract—Generative world modeling from large-scale visual data offers a promising path toward scalable robot manipulation. This calls for an intermediate representation that can be supervised at scale and remains actionable for downstream control. We propose 3D full-object flow, an object-centric motion representation that assigns dense 3D trajectories to points across the complete surface of a target object, including its occluded regions. By extending visible-surface object flow to the full object geometry, this representation makes complete object motion explicit and applies uniformly to both rigid and deformable objects. To realize the potential of this representation at scale, we introduce a 1210-hour full-object flow dataset that addresses the scarcity of large-scale supervision for complete 3D object motion. The dataset integrates diverse real-world videos and physical simulations, with a scalable annotation pipeline for recovering full-object motion from real-world observations and quality metrics for systematically assessing annotation quality. We then train a world model that takes an image and a task description as input and predicts dense full-object flow over the reconstructed object geometry. Experiments show reliable annotation quality, data-scalable flow prediction, and effective downstream manipulation across novel scenes and robot embodiments. Project website with additional results and videos: <https://nus-lins-lab.github.io/ObjectWorld/>.

[†] denotes the corresponding author.

¹School of Computing, National University of Singapore; ²RoboScience; ³NUS Guangzhou Research Translation and Innovation Institute

I. INTRODUCTION

Recent progress in robot foundation models and vision-language-action (VLA) models suggests that robotic manipulation can benefit substantially from scale [1]–[3]. However, action-labeled robot data remain costly to collect, embodiment-dependent, and far less abundant than visual data [3], [4]. Generative world models therefore offer a promising path beyond direct action supervision by learning how objects and scenes evolve from visual experience [5]–[9]. Connecting these predictions to robot execution requires an intermediate representation that is both scalable to learn and actionable for downstream control [9], [10].

Prior representations each capture useful structure but expose trade-offs. Pixel-space predictions, such as goal images [11] or videos [8], preserve rich visual information, but they are indirect for control and typically require an additional grounding step. Sparse keypoint or point-based interfaces represent manipulation objectives using landmarks or affordance points [12], [13], but may miss the complete object motion. 6-DoF object poses offer metrically grounded state descriptions, but they rely on rigid-body assumptions and cannot capture non-rigid deformations [14], [15]. Recent flow-based interfaces use object flow as an intermediate

representation for manipulation [16]–[18]. However, they mainly model motion over visible or tracked object regions, leaving unobserved surfaces outside the motion field. We refer to object flow over visible regions as partial object flow.

To make object motion complete and actionable, we introduce 3D full-object flow, which specifies dense trajectories over the complete target-object surface as an explicit motion target for robot execution. Utilizing the reconstructed geometry of a target object, this representation maps dense trajectories over the entire object surface across time. It makes complete object motion explicit in a unified 3D geometry, while remaining grounded in object kinematics rather than robot-specific actions. The resulting dense correspondence can describe rigid motion and non-rigid deformation, and provides an object-level motion target that downstream controllers can optimize against.

A key obstacle to adopting and scaling this complete representation is the scarcity of high-quality training data [19]–[21]. While numerous large-scale human video datasets and robotic trajectory collections exist [3], [4], [22], [23], their supervision signals are predominantly limited to semantic action labels, sparse 2D tracking, or robot-specific proprioception. They lack the dense, unified 3D full-object flow annotations required to supervise a generalizable object-centric world model across diverse tasks and scenes.

In this work, we address both the representation choice and the supervision bottleneck. We construct a 1210.1-hour 3D full-object flow dataset from diverse real-world videos and physical simulation. To build the real-world portion at scale, we introduce an automatic annotation pipeline that converts in-the-wild manipulation videos into 3D full-object flow, and we systematically assess the quality of the resulting annotations. On top of this dataset, we develop an object-centric world model framework that takes an image and a task description, reconstructs the initial geometry of the target object, and generate dense trajectories over the full object surface with a flow generator. The generated full-object flow is then consumed by a lightweight optimization-based execution module that translates object motion into end-effector trajectories. Experiments validate the full pipeline from annotation to downstream execution, demonstrating reliable flow supervision, improved prediction through data scaling and model design, and effective robot manipulation with transfer across environments and embodiments.

Contributions. (i) We formulate 3D full-object flow as an object-centric interface between generative world models and downstream robot execution. (ii) We provide a 1210.1-hour 3D full-object flow dataset built from diverse real-world videos and physical simulation, together with a scalable annotation pipeline and quality assessment for the generated annotations. (iii) We develop an object-centric world model framework that predicts dense full-object trajectories from image-and-language inputs. (iv) We conduct extensive experiments across annotation, data-scale, prediction, and execution, showing effective robot manipulation and generalization to novel scenes and embodiments.

II. RELATED WORKS

A. World Models for Robotic Manipulation

World models have been increasingly used in robotic manipulation to predict future states or intermediate representations for planning and control [5], [9], [10]. A common strategy is to use intermediate representations less tied to a specific action space or embodiment, such as future images [11], videos [8], or keypoints [12], [13]. These abstractions can improve transfer across tasks and embodiments, but are often indirect for control or sparse in object coverage.

Recent work has begun to model object motion more directly. Pose-based methods use 6-DoF object poses to represent future object states or task goals [14], [15], [24], providing metric structure for rigid-object manipulation but not naturally capturing non-rigid deformation. Im2Flow2Act [16] uses 2D object flow over tracked image points, Dream2Flow [17] and NovaFlow [18] derive object-centric 3D flow from masked tracked points in generated videos, and PointWorld [6] predicts action-conditioned full-scene 3D point flow from RGB-D observations and robot actions. These works demonstrate the value of flow-based interfaces, while their prediction targets are typically defined on image tracks, masked visible regions, or partially observed scene geometry. Our work predicts full-surface 3D trajectories over a reconstructed complete target object for downstream optimization.

B. Datasets and Scalable Supervision for 3D Object Flow

Scalable supervised learning of dense object-centric 3D motion prediction benefits from large-scale 3D motion annotations, yet such supervision remains limited in real-world data [6], [19], [21]. Large video datasets [4], [22], [25], [26] provide coverage of human-object interaction, but their annotations do not provide dense 3D trajectories of manipulated objects. Datasets with richer 3D annotations, such as HOI4D [27], DexYCB [28], ARCTIC [29], and HOT3D [30], provide object poses, meshes, contacts, or tracking signals. However, their reliance on curated capture setups, markers, or labor-intensive annotation limits their scalability for dense object-flow supervision. Large robot datasets, such as Open X-Embodiment [3], DROID [23], and AgiBot World [31], further scale real-world manipulation data, but they primarily provide robot observation-action trajectories rather than unified 3D full object flow annotations. To address this gap, we build a scalable annotation pipeline that converts heterogeneous real-world videos into 3D full object flow supervision, emphasizing dense target object motion rather than action labels, sparse poses, or robot-specific trajectories.

III. METHOD

We first formulate object-centric world modeling and flow-conditioned robot execution in Sec. III-A. We then describe the annotation pipeline and dataset construction in Sec. III-B, the world model architecture and training strategy in Sec. III-C, and the downstream action optimization in Sec. III-D.

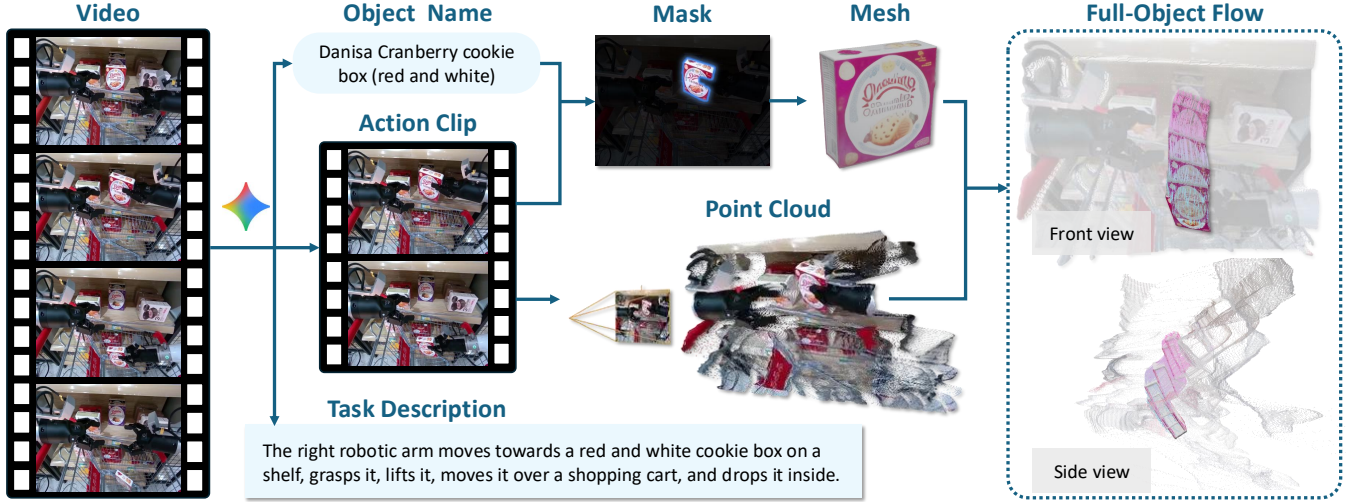


Fig. 2: **Automatic data annotation pipeline.** Given an in-the-wild manipulation video, a VLM identifies the target object, describes the task, and segments the video into atomic action clips. For each clip, our pipeline extracts the target-object masks, reconstructs the complete object mesh and the point clouds, and aligns the mesh with the point clouds across frames to recover the object motion. The motion is propagated over the complete object surface to produce 3D full-object flow.

A. Problem Formulation

Given an initial RGBD observation $\mathcal{I}$ and a natural language instruction $\mathcal{L}$, our world model predicts a 3D full-object flow $\hat{\mathcal{P}}_{1:T}^{\text{obj}} = \{\hat{\mathbf{p}}_{i,t} \in \mathbb{R}^3\}_{i=1,t=1}^{N^{\text{obj}},T}$, where $N^{\text{obj}} = 512$ is the number of sampled object points and $T = 48$ is the prediction horizon. Each $\hat{\mathbf{p}}_{i,t}$ denotes the predicted 3D position of object point i at time step t . This representation specifies the desired object motion for accomplishing the task, and serves as an interface between high-level task prediction and low-level robot execution. Conditioned on the predicted full-object flow, the robot executes the task by solving a flow-conditioned action optimization problem:

$$\tau^* = \arg \min_{\tau \in \mathcal{A}} \mathcal{D}(\Phi(\mathcal{I}, \tau), \hat{\mathcal{P}}_{1:T}^{\text{obj}}) + \lambda \mathcal{R}(\tau),$$

where τ denotes the robot action variable, instantiated as an end-effector trajectory in our implementation; $\mathcal{A}$ is the feasible trajectory space; $\Phi(\mathcal{I}, \tau)$ denotes the object motion induced by executing τ from the initial observation; $\mathcal{D}$ measures its discrepancy to the predicted full-object flow; and $\mathcal{R}$ regularizes action feasibility and smoothness.

B. Dataset Curation and Pipeline

1) *Annotation of Real-World Videos:* To enable scalable supervision, we build an annotation pipeline that converts heterogeneous in-the-wild videos into 3D full-object flow, as illustrated in Fig. 2. The pipeline does not require robot actions, depth measurements, calibrated cameras, or pre-scanned object meshes. It first uses Gemini 2.5 Pro [32] as a Vision-Language Model (VLM) to identify the target object, generate a task description, and segment each raw video into atomic manipulation clips, each corresponding to a single contact-and-separation interaction between the actor and the target object. For each clip, we use Track4World [33] with its Depth Anything 3 (DA3) backbone [34] to estimate camera poses and reconstruct per-frame point clouds in a common coordinate system anchored to the first-frame camera.

We then localize the target object using VLM-based object proposals followed by SeC [35] for mask propagation. Given the propagated target masks, SAM 3D [36] reconstructs the complete object mesh, Any6D [37] estimates its initial pose, and FoundationPose [38] tracks its pose throughout the clip. To obtain full-object flow, we uniformly sample points from the reconstructed mesh in its canonical object frame and, at each timestep, transform them by composing the tracked object pose with the corresponding camera-to-reference transformation recovered by Track4World. The resulting dense trajectories, expressed in the first-frame camera coordinate system, constitute the full-object flow $\mathcal{P}_{1:T}^{\text{obj}}$.

2) *Data Filtering:* To filter low-quality annotations, we evaluate the spatial and temporal consistency of the estimated object motion. At each frame, we project the reconstructed object under its estimated pose into the image plane and compute the IoU between the projected object region and the propagated target mask. We aggregate the IoU scores over the clip to identify annotations with unstable pose tracking.

We manually curate a validation set of expert-labeled examples to calibrate the filtering criterion. We evaluate the consistency metric on this set and select a threshold that achieves high precision while retaining sufficient coverage according to the expert labels. We then apply the threshold across the dataset to remove low-quality annotations.

3) *Dataset Composition:* Our dataset comprises 1,210.1 hours of manipulation data measured at a standardized 16 FPS: 454.5 hours of simulated trajectories and 755.6 hours of automatically annotated real-world videos.

For the simulation subset, we aggregate 198,947 trajectories spanning 251 tasks from seven datasets: ManiSkill 3 [41], MetaFold [42], Meta-World [43], MIKASA [44], RL Bench [45], RoboCasa [46], and RoboTwin 2.0 [47]. MetaFold contributes deformable-object manipulation data.

For real-world data, we collect videos with diverse viewpoints and embodiments from three sources:

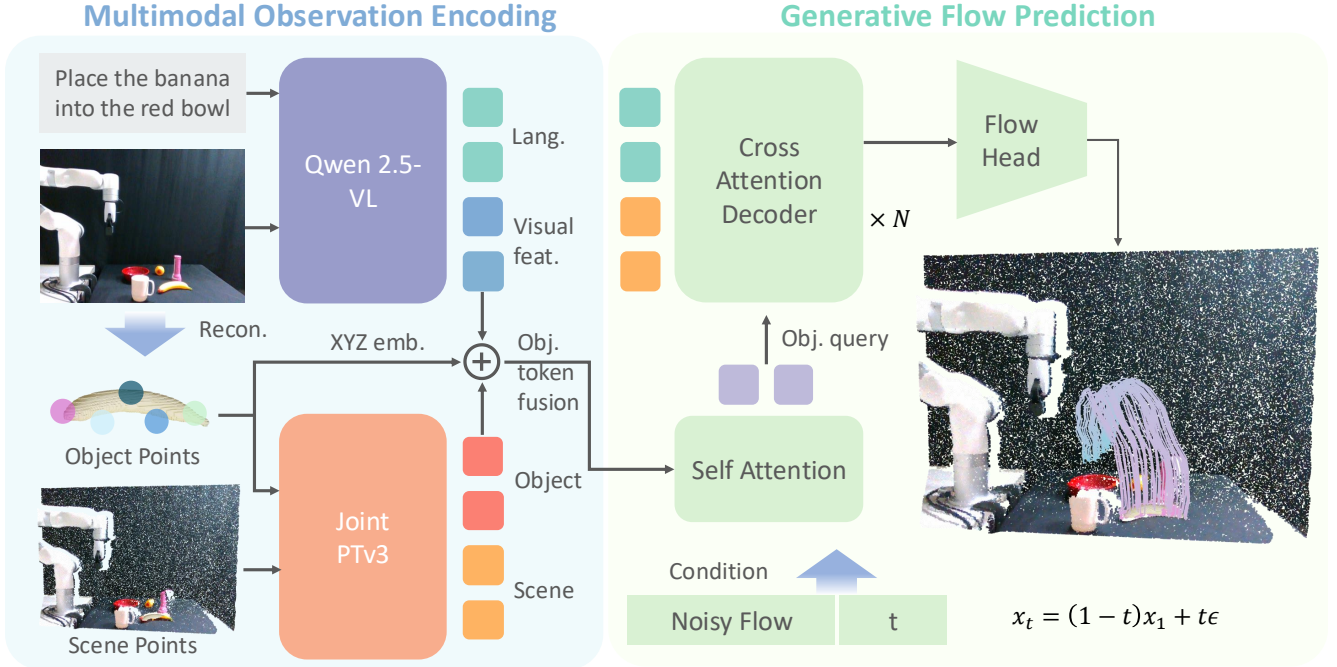


Fig. 3: **Object-centric full-object flow prediction framework.** Given an initial RGB-D observation and task instruction, Qwen2.5-VL [39] extracts task-conditioned visual-language features, while a joint PointTransformer V3 [40] (PTv3) encodes reconstructed object and scene geometry. These features are fused into object tokens and refined by self-attention. The object queries attend to scene and language context through a cross-attention decoder, conditioned on noisy flow and time t , before a flow head predicts dense trajectories over the complete object surface.

robot demonstrations from AgiBot World [31], third-person human demonstrations from Something-Something V2 [25], and egocentric human demonstrations from EPIC-KITCHENS [22]. After automatic annotation and quality filtering, we retain 289,976 clips totaling 755.6 hours.

C. Full-Object Flow Prediction

Given an initial observation and a task instruction, we predict the future motion of the complete target-object surface, as illustrated in Fig. 3. Let $\mathbf{P}^o \in \mathbb{R}^{N_o \times 3}$ denote the initial object surface and $\mathbf{P}^s \in \mathbb{R}^{N_s \times 3}$ the scene geometry. Our predictor outputs a dense displacement trajectory

$$\mathbf{F} = \{\Delta \mathbf{P}_t^o\}_{t=1}^T \in \mathbb{R}^{T \times N_o \times 3}, \quad (1)$$

where each displacement is defined with respect to the initial object surface. This object-centric representation describes how the manipulated object evolves independently of the robot embodiment.

a) Initial 3D observation construction.: We construct an object-centric 3D observation from the first RGB-D frame. We first obtain the target-object mask using SAM2 [48] and reconstruct its complete geometry with SAM3D [36]. Any6D [37] estimates the initial object pose, which places the reconstructed mesh into the observed scene. We then sample points on the transformed mesh surface to obtain the complete object point set $\mathbf{P}^o$. In parallel, the scene point cloud $\mathbf{P}^s$ is obtained by back-projecting the observed depth map using the camera intrinsics. Thus, the model receives both a complete object-centric geometric representation and the observed scene geometry from the initial frame.

b) Multimodal observation encoding.: We encode semantic and geometric information through complementary visual-language and 3D pathways. A vision-language encoder processes the initial RGB image together with the task instruction and produces task-conditioned visual features and language tokens. For each object point, we project its 3D coordinate into the image and sample the corresponding visual feature, providing a point-wise semantic representation.

In parallel, we jointly encode the object and scene point clouds using Point Transformer V3 (PTv3) [40]. Joint 3D encoding allows the object representation to incorporate its surrounding geometric context. For each object point $\mathbf{p}_i^o$, we construct an object token by combining its spatial embedding, point-wise visual-language feature, and 3D feature:

$$\mathbf{h}_i^o = \phi_{xyz}(\mathbf{p}_i^o) + \phi_{vlm}(\mathbf{f}_i^{vlm}) + \phi_{3D}(\mathbf{f}_i^{3D}). \quad (2)$$

The scene-side PTv3 features are retained as contextual scene tokens, while the language tokens preserve the task-level semantic information.

c) Generative full-object flow prediction.: We formulate future-flow prediction using an x_1 -prediction parameterization over a linear noising path. Let $\mathbf{x}_1$ denote the full-object trajectory and $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote Gaussian noise. During training, we sample $t \sim \mathcal{U}(0, 1)$ and construct

$$\mathbf{x}_t = (1 - t)\mathbf{x}_1 + t\mathbf{x}_0. \quad (3)$$

The noisy trajectory $\mathbf{x}_t$ is projected into the object tokens, while the timestep t modulates them through adaptive normalization. The conditioned object tokens are refined by self-attention and then used as queries in a cross-attention

decoder, whose keys and values consist of the scene and language tokens. This allows each object point to reason jointly about the task semantics, surrounding geometry, and the motion of the complete object surface.

The network directly predicts the clean trajectory,

$$\hat{\mathbf{x}}_1 = f_\theta(\mathbf{x}_t, t, \mathbf{H}^o, \mathbf{H}^s, \mathbf{H}^l), \quad (4)$$

where $\mathbf{H}^o$, $\mathbf{H}^s$, and $\mathbf{H}^l$ denote the object, scene, and language representations, respectively. All T future frames are predicted jointly. We optimize

$$\mathcal{L}_{\text{flow}} = \|\hat{\mathbf{x}}_1 - \mathbf{x}_1\|_2^2. \quad (5)$$

At inference time, we initialize the trajectory from Gaussian noise and apply the same x_1 -prediction update for one or more refinement steps to obtain the full-object trajectory.

d) Training with Heterogeneous Data: We train the flow predictor on heterogeneous simulated and real-world data, which introduces inconsistencies in coordinate orientation and temporal scale. To standardize orientation, we estimate gravity with GeoCalib [49] and rotate each scene such that the z -axis points opposite to gravity, applying the same transformation to both 3D points and object-flow trajectories. This normalization provides a consistent gravity-aligned coordinate frame across datasets, which is important because manipulation dynamics depend on gravity [24]. To normalize temporal scale, we uniformly resample each action segment into T frames, such that object flow represents spatial evolution over normalized action progress rather than absolute execution time.

D. Flow-Conditioned Action Optimization

We execute the predicted full-object flow with a lightweight local patch fitting optimizer. A preceding grasp module provides the initial manipulator pose $\mathbf{X}_1^m = (\mathbf{R}_1^m, \mathbf{t}_1^m)$ and the contact region. We use sample-based grasping with a parallel gripper and T(R,O) Grasp [50] for the dexterous hand. Let $\mathcal{C} \subseteq \{1, \dots, N^{\text{obj}}\}$ denote a local contact patch. We express its points in the initial manipulator frame as

$$\mathbf{y}_i = (\mathbf{R}_1^m)^\top (\mathbf{p}_{i,1} - \mathbf{t}_1^m), \quad i \in \mathcal{C}. \quad (6)$$

The local coordinates $\mathbf{y}_i$ are kept fixed when recovering the trajectory. This imposes only a local rigidity assumption around the contact patch, rather than a global rigidity assumption on the object; the predicted full-object flow can still represent non-rigid motion over the rest of the surface.

At each time step, we recover the manipulator pose by robustly aligning this local patch with its predicted positions:

$$(\mathbf{R}_t^m, \mathbf{t}_t^m) = \arg \min_{\mathbf{R} \in SO(3), \mathbf{t} \in \mathbb{R}^3} \sum_{i \in \mathcal{C}} \rho_\delta (\|\mathbf{R}\mathbf{y}_i + \mathbf{t} - \hat{\mathbf{p}}_{i,t}\|_2). \quad (7)$$

The resulting poses form the execution trajectory $\tau = \{\mathbf{X}_t^m\}_{t=1}^T$. We solve the Huber-robust objective using IRLS, with each subproblem given by weighted Kabsch alignment.

IV. EXPERIMENTS

Our experiments address the following key questions:

- Q1:** How reliable are the full-object flow annotations produced and retained by our large-scale pipeline?
- Q2:** How accurately does our model predict full-object flow?
- Q3:** How do data scale and model design affect prediction?
- Q4:** Can full-object flow serve as an effective motion target for downstream robot manipulation?
- Q5:** How well does our system generalize across different environments and robot embodiments?

TABLE I: Experiment map. A: annotation validation; B: flow prediction and scaling; C: robot execution and generalization.

ID	Study	Supports
A1	Pipeline validation against simulator ground truth	Q1
A2	Filter-gate validation against expert labels	Q1
A3	Full-object vs. partial-object flow annotation	Q1
B1	Flow predictor comparison with baselines	Q2
B2	Robustness of flow predictor	Q2
B3	Data scaling experiment	Q3
B4	Task-language ablation study	Q3
C1	Simulation execution with predicted flow	Q4, Q5
C2	Real-world execution with predicted flow	Q4, Q5

A. Experimental Setup

Evaluation strategy. We organize the evaluation around the five questions above and group the studies into three families: annotation validation (Q1), flow prediction (Q2), data scaling and ablation (Q3), and robot execution and generalization (Q4–Q5), as summarized in Table I.

Platforms and embodiments. We evaluate in-domain simulation performance in ManiSkill 3 [41] and out-of-domain transfer in LIBERO [51]. Real-world evaluation covers two settings, with parallel-gripper and dexterous-hand systems performing different tasks in their respective environments.

Training data. Using the filtering protocol validated in Fig. 4, we retain 288.7 hours of real-world videos as the Golden subset. Unless otherwise specified, our method and the learning-based baselines are trained on all simulation data together with this Golden subset.

B. Annotation Pipeline Validation (Q1)

We evaluate the reliability of our annotation pipeline through three complementary experiments: validation against simulator ground truth, validation of the filtering gates against expert judgments, and a comparison between full-object and partial-object flow annotations.

Validation against simulator ground truth. We evaluate each annotation stage on 99 eligible 3-second RoboTwin clips. The pipeline successfully processes 97 clips, yielding a 98.0% annotation rate. As shown in Table II, target localization and mask propagation achieve median IoUs of 0.908 and 0.873, respectively. Complete-object reconstruction obtains a Chamfer distance of 2.61% D , with a hidden-surface error of 7.28% D under visible-only alignment. After pose estimation and tracking, the recovered surface motion has a trajectory error of 11.50% D , and the final dense full-object flow achieves an EPE of 3.71 cm (26.11% D). These results

TABLE II: **A1. Ground-truth validation and runtime of the annotation pipeline on simulation.** Quality metrics are medians over 97 successfully annotated RoboTwin clips. D denotes the ground-truth object diameter. Times are measured on standardized 3-second clips. [†]Amortized time per call; [‡]CPU flow construction.

Module / Stage	Evaluated Output	Protocol	Metric	Median	Time (s)
Gemini 2.5 Pro (VLM)	Localization	Against GT box	Box IoU $\uparrow$	0.908	12.04 [†]
SeC	Mask propagation	Against GT masks	Mask IoU $\uparrow$	0.873	11.20
Track4World (DA3)	Relative depth	Median-scale aligned	AbsRel $\downarrow$ / $\delta < 1.25 \uparrow$	0.089 / 0.826	8.25
SAM 3D Objects	Complete geometry	Shape aligned	CD/ $D \downarrow$	2.61%	7.41
	Hidden geometry	Visible-only Sim(3)	Vis./hid. surface error/ $D \downarrow$	4.22/7.28%	
	Hidden coverage	Visible-only Sim(3)	Recall@5% $D \uparrow$	0.462	
Any6D + FoundationPose	Surface motion	Initial-frame Sim(3)	Surface-trajectory error/ $D \downarrow$	11.50%	5.17+3.53
3D full-object flow	Dense trajectories	Frozen correspondence	EPE [cm] / EPE/ $D \downarrow$	3.71 / 26.11%	0.19 [‡]
Reprojection-IoU filter	Quality selection	Golden vs. rejected	Frozen-flow EPE/ $D \downarrow$	16.29/35.49%	–

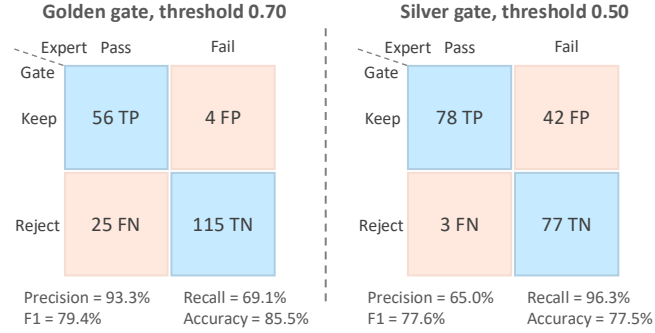


Fig. 4: **A2. Expert validation of the filtering gates.** Confusion matrices for the Golden and Silver gates on 200 expert-labeled real-world clips.

TABLE III: **A3. Full-object versus partial motion supervision.** All results are medians. EPE is evaluated on identical GT material points; Coverage@5% D is measured over the complete GT surface, with unsuccessful cases counted as 0.

Representation	EPE/ $D \downarrow$	Surface Cov. $\uparrow$	Hidden Cov. $\uparrow$
Track4World partial	37.22%	28.4%	9.8%
Full-object flow	30.74%	45.8%	26.6%

indicate that the successive pipeline stages recover consistent object geometry and motion from monocular videos. The local GPU stages require approximately 35.6 seconds per clip and can be parallelized across clips.

Filtering against expert judgments. We examine whether the filtering criterion can distinguish reliable annotations from failure cases. Under simulator ground truth, annotations retained by the Golden gate have a substantially lower median flow error than rejected annotations (16.29% D versus 35.49% D). We further evaluate the gates on 200 expert-labeled real-world clips. As shown in Fig. 4, the Golden gate achieves 93.3% precision, while the Silver gate achieves 96.3% recall. The two thresholds provide complementary operating points: the Golden gate prioritizes annotation quality, whereas the Silver gate retains broader coverage.

Full-object versus partial motion supervision. In Table III, we compare our full-object annotations with Track4World trajectories anchored to the initially visible object surface. On identical GT material points, full-object flow reduces median

EPE from 37.22% D to 30.74% D . It also increases complete-surface and hidden-surface Coverage@5% D from 28.4% to 45.8% and from 9.8% to 26.6%, respectively. These results show that full-object flow provides more accurate median motion while extending valid supervision beyond the initially observed surface.

C. Full-Object Flow Prediction (Q2)

Flow-prediction baseline comparison. Table IV(a) compares our model with the partial-flow prediction methods TraceGen and im2Flow2Act, as well as ObjectForesight, which models rigid-object motion. In contrast, our model predicts dense trajectories over the complete object surface and accommodates both rigid and non-rigid motion. It achieves the best results across all all-frame and final-frame metrics, with EPEs of 6.22 cm and 11.18 cm, respectively. This advantage shows that our model more accurately captures intermediate motion and long-horizon displacement.

Robustness to object-geometry noise. We further perturb the reconstructed object points with i.i.d. Gaussian noise while keeping the scene geometry and target flow unchanged. Table IV(b) shows negligible degradation for noise levels up to 2 cm. Even with 5 cm noise, final-frame EPE remains 11.61 cm and Acc@10 decreases only from 60.1% to 58.3%. Thus, the predictor is robust to moderate errors in the reconstructed object geometry.

D. Data Scaling and Ablation Study (Q3)

Data scaling. We align the number of training epochs across all data configurations and evaluate their performance.

As shown in Fig. 6, prediction error generally decreases as either simulated or real-world training data increases. Simulated data provides the larger overall gain, while real-world data offers complementary supervision for motions and observations encountered in real environments. Overall,

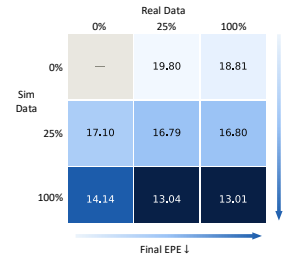


Fig. 6: **B3. Data scaling.**

scaling either data source improves prediction accuracy, and combining the two yields the strongest performance.

TABLE IV: **Flow predictor evaluation.** EPE is reported in centimeters, and $\text{Acc}@ \tau$ denotes per-point accuracy (%). Subscript T denotes the final frame. $^\dagger \text{im2Flow2Act}$ reports lateral 2D-flow error.

(a) B1. Comparison with flow-prediction baselines.							(b) B2. Robustness to object-geometry noise.				
Method	All Frames			Final Frame			σ (cm)	EPE _{all} ↓	EPE _T ↓	Acc@5 _T ↑	Acc@10 _T ↑
	EPE↓	Acc@5↑	Acc@10↑	EPE↓	Acc@5↑	Acc@10↑					
ObjectForesight	10.24	48.9	64.3	21.18	9.9	25.9	0	6.22	11.17	28.6	60.1
TraceGen	12.97	42.2	54.2	26.43	3.3	11.1	0.5	6.23	11.17	28.8	60.0
im2Flow2Act [†]	12.58	36.7	63.2	20.88	18.8	38.0	1	6.23	11.17	28.9	60.2
Ours	6.22	61.0	79.3	11.18	28.6	60.1	2	6.23	11.22	28.8	60.0
							5	6.34	11.61	26.6	58.3

Fig. 5: **Qualitative full-object flow predictions.** Given an initial RGB observation and task instruction (top), our model predicts dense 3D trajectories over the complete target-object surface (bottom) across real-world and simulated tasks.

TABLE V: **B4. Task-language ablation.** Final-frame EPE.

Caption	AgiBot	SthV2	Overall
Placeholder	16.39	17.48	13.76
Task-specific	14.86	15.27	13.01
Δ	-1.53	-2.21	-0.75

TABLE VI: **C1. Simulation execution results.**

Method	ManiSkill	LIBERO	
	PushCube	Insert	Stack-Arrange
TraceGen	5/10	7/10	7/10
ObjectForesight	6/10	7/10	6/10
Ours	10/10	9/10	10/10

Effect of task-language conditioning. Table V examines whether task-specific language provides useful motion cues. Replacing placeholder captions with task descriptions reduces overall final-frame EPE from 13.76 to 13.01 cm. The largest gains occur on AgiBot and SthV2, where EPE decreases by 1.53 and 2.21 cm, respectively. These results confirm that task language provides meaningful conditioning for full-object motion prediction, particularly on heterogeneous real-world data.

E. Full-Object Flow as an Execution Interface (Q4)

Tables VI and VII report the simulation and real-world execution results, respectively. All methods share the same grasping, flow-to-action optimization, and low-level execution pipeline, with the pre-grasp phase standardized. We assess post-grasp execution using its deviation from the ground-truth trajectory, providing a finer-grained measure than a binary task outcome; executions within a fixed tolerance are considered successful. Under this common pipeline,

TABLE VII: **C2. Real-world execution results.**

Method	Banana	Water	Flashlight	Overall
Dream2Flow	2/5	3/5	5/5	10/15
Ours	4/5	4/5	5/5	13/15

our method generally achieves more accurate downstream execution than the baselines. More broadly, converting predicted full-object flow into manipulator trajectories supports its viability as an execution interface.

F. Environment and Embodiment Generalization (Q5)

We evaluate transfer across environments and robot embodiments. Our method and the baselines are tested in the in-domain ManiSkill environment and the held-out LIBERO environment using different scenes and grippers, with results reported in Table VI. In the real world, we quantitatively evaluate a parallel gripper on two pick-and-place tasks involving a banana and a flashlight, and a water-pouring task (Table VII), while presenting dexterous-hand results qualitatively in Fig. 5. Together, these results show that the object-flow interface can accommodate changes in both environment and embodiment.

V. CONCLUSION

We introduced 3D full-object flow as an object-centric interface between predictive world models and downstream robot execution. By assigning dense trajectories to the complete target-object surface, this representation captures both visible and occluded object motion while remaining independent of a specific robot embodiment. We further built a large-scale full-object flow dataset from real-world videos and simulation, and trained a world model that

predicts complete object motion from image-and-language inputs. Experiments demonstrate reliable annotation quality, improved flow prediction with scalable data, and effective downstream manipulation across simulation and real-world settings. These results suggest that full-object 3D motion prediction is a promising foundation for scalable and transferable robotic world models.

ACKNOWLEDGMENT

This work was supported by the NUS Guangzhou Research Translation and Innovation Institute (GRTII)-Affiliated PhD Scholarship Program. We thank Zhehao Zhou, Jiongxu Chen, Zhuoman Liu, Yiyao Liu, Xinrui Qi, Minwen Liao, Yichen Guo, and the RoboScience data and simulation team for their valuable support and assistance.

REFERENCES

- [1] M. J. Kim, K. Pertsch, *et al.*, “Openvla: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.
- [2] B. Zitkovich, T. Yu, S. Xu, *et al.*, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.
- [3] A. O’Neill, A. Rehman, A. Maddukuri, *et al.*, “Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0,” in *ICRA*, 2024.
- [4] K. Grauman, A. Westbury, E. Byrne, *et al.*, “Ego4d: Around the world in 3,000 hours of egocentric video,” in *CVPR*, 2022.
- [5] C. Finn and S. Levine, “Deep visual foresight for planning robot motion,” in *ICRA*, 2017.
- [6] W. Huang, Y.-W. Chao, A. Mousavian, *et al.*, “Pointworld: Scaling 3d world models for in-the-wild robotic manipulation,” *CVPR*, 2026.
- [7] C. Gao, H. Zhang, *et al.*, “Flip: Flow-centric generative planning as general-purpose manipulation world model,” in *ICLR*, 2025.
- [8] J. Wu, S. Yin, N. Feng, X. He, D. Li, J. Hao, and M. Long, “ivideogpt: Interactive videogpts are scalable world models,” *NeurIPS*, 2024.
- [9] B. Hou, G. Li, J. Jia, *et al.*, “World model for robot learning: A comprehensive survey,” *arXiv preprint arXiv:2605.00080*, 2026.
- [10] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine, “Visual foresight: Model-based deep reinforcement learning for vision-based robotic control,” *arXiv preprint arXiv:1812.00568*, 2018.
- [11] K. Black, M. Nakamoto, *et al.*, “Zero-shot robotic manipulation with pre-trained image-editing diffusion models,” in *ICLR*, 2024.
- [12] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, “Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation,” *arXiv preprint arXiv:2409.01652*, 2024.
- [13] F. Liu, K. Fang, P. Abbeel, *et al.*, “Moka: Open-vocabulary robotic manipulation through mark-based visual prompting,” in *RSS*, 2024.
- [14] R. Soraki, H. Bharadhwaj, A. Farhadi, and R. Mottaghi, “Objectfore-sight: Predicting future 3d object trajectories from human videos,” *arXiv preprint arXiv:2601.05237*, 2026.
- [15] H. Chen, J. Guo, B. Wang, *et al.*, “Goal-vla: Image-generative vlms as object-centric world models empowering zero-shot robot manipulation,” *arXiv preprint arXiv:2506.23919*, 2025.
- [16] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, “Flow as the cross-domain manipulation interface,” in *CoRL*, 2024.
- [17] K. Dharmarajan *et al.*, “Dream2flow: Bridging video generation and open-world manipulation with 3d object flow,” *ICRA*, 2026.
- [18] H. Li, L. Sun, Y. Hu, D. Ta, J. Barry, G. Konidaris, and J. Fu, “Novaflow: Zero-shot manipulation via actionable flow from generated videos,” *arXiv preprint arXiv:2510.08568*, 2025.
- [19] R. Firoozi, J. Tucker, S. Tian, *et al.*, “Foundation models in robotics: Applications, challenges, and the future,” *IJRR*, 2025.
- [20] C. Gao, Z. Liu, *et al.*, “Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models,” *NeurIPS*, 2026.
- [21] Z. Wang, B. Wang, H. Zhang, *et al.*, “Vision-language-action in robotics: A survey of datasets, benchmarks, and data engines,” *arXiv preprint arXiv:2604.23001*, 2026.
- [22] D. Damen, H. Doughty, G. M. Farinella, *et al.*, “The epic-kitchens dataset: Collection, challenges and baselines,” *TPAMI*, 2020.
- [23] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, *et al.*, “Droid: A large-scale in-the-wild robot manipulation dataset,” *RSS*, 2024.
- [24] H. Chen, T. Dong, T. Wu, *et al.*, “Dexterous manipulation policies from rgb human videos via 3d hand-object trajectory reconstruction,” *arXiv preprint arXiv:2602.09013*, 2026.
- [25] R. Goyal, S. Ebrahimi Kahou, V. Michalski, *et al.*, “The” something something” video database for learning and evaluating visual common sense,” in *ICCV*, 2017.
- [26] F. Mahdisoltani, G. Berger, W. Gharbieh, D. Fleet, and R. Memisevic, “On the effectiveness of task granularity for transfer learning,” *arXiv preprint arXiv:1804.09235*, 2018.
- [27] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, “Hoi4d: A 4d egocentric dataset for category-level human-object interaction,” in *CVPR*, 2022.
- [28] Y.-W. Chao, W. Yang, Y. Xiang, *et al.*, “Dexycb: A benchmark for capturing hand grasping of objects,” in *CVPR*, 2021.
- [29] Z. Fan, O. Taheri, D. Tzionas, *et al.*, “Arctic: A dataset for dexterous bimanual hand-object manipulation,” in *CVPR*, 2023.
- [30] P. Banerjee, S. Shkodrani, P. Moulon, *et al.*, “HOT3D: Hand and object tracking in 3D from egocentric multi-view videos,” *CVPR*, 2025.
- [31] Q. Bu, J. Cai, L. Chen, *et al.*, “Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems,” *arXiv preprint arXiv:2503.06669*, 2025.
- [32] G. Comanici, E. Bieber, *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.
- [33] J. Lu, J. Xu, *et al.*, “Track4world: Feedforward world-centric dense 3d tracking of all pixels,” *arXiv preprint arXiv:2603.02573*, 2026.
- [34] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, “Depth anything 3: Recovering the visual space from any views,” *arXiv preprint arXiv:2511.10647*, 2025.
- [35] Z. Zhang, S. Ding, X. Dong, *et al.*, “Advancing complex video object segmentation via progressive concept construction,” *arXiv preprint arXiv:2507.15852*, 2025.
- [36] X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, *et al.*, “Sam 3d: 3dfy anything in images,” *arXiv preprint arXiv:2511.16624*, 2025.
- [37] T. Lee, B. Wen, M. Kang, *et al.*, “Any6d: Model-free 6d pose estimation of novel objects,” in *CVPR*, 2025.
- [38] B. Wen, W. Yang, J. Kautz, *et al.*, “Foundationpose: Unified 6d pose estimation and tracking of novel objects,” in *CVPR*, 2024.
- [39] S. Bai, K. Chen, X. Liu, *et al.*, “Qwen2.5-vl technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.13923>
- [40] X. Wu, L. Jiang, P.-S. Wang, *et al.*, “Point transformer v3: Simpler faster stronger,” in *CVPR*, 2024.
- [41] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, *et al.*, “Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai,” *arXiv preprint arXiv:2410.00425*, 2024.
- [42] H. Chen, J. Li, R. Wu, *et al.*, “Metafold: Language-guided multi-category garment folding framework via trajectory generation and foundation model,” in *IROS*, 2025.
- [43] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning,” in *CoRL*. PMLR, 2020.
- [44] E. Cherepanov, N. Kachaev, A. K. Kovalev, and A. I. Panov, “Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning,” *arXiv preprint arXiv:2502.10550*, 2025.
- [45] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, “Rlbench: The robot learning benchmark & learning environment,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.
- [46] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, *et al.*, “Robocasa: Large-scale simulation of everyday tasks for generalist robots,” *arXiv preprint arXiv:2406.02523*, 2024.
- [47] T. Chen, Z. Chen, *et al.*, “Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation,” *arXiv preprint arXiv:2506.18088*, 2025.
- [48] N. Ravi, V. Gabeur, Y.-T. Hu, *et al.*, “Sam 2: Segment anything in images and videos,” in *ICLR*, 2025.
- [49] A. Veicht, P.-E. Sarlin, P. Lindenberger, *et al.*, “Geocalib: Learning single-image calibration with geometric optimization,” in *ECCV*, 2024.
- [50] X. Fei, Z. Xu, H. Fang, *et al.*, “T (r, o) grasp: Efficient graph diffusion of robot-object spatial transformation for cross-embodiment dexterous grasping,” *arXiv preprint arXiv:2510.12724*, 2025.
- [51] B. Liu, Y. Zhu, C. Gao, *et al.*, “Libero: Benchmarking knowledge transfer for lifelong robot learning,” *NeurIPS*, 2023.