

F.R.A.N.T.Z.: 3D Scene Reconstruction and Hand-Trajectory Extraction from Monocular Video via Training-Light Compositional AI

Frantz Lohier, PhD
ShedInnovations LLC
California, USA
flohier@gmail.com

Abstract—We present F.R.A.N.T.Z., an open, modular, inspectable perception pipeline that composes pre-trained AI modules—MiDaS/DAV2 for monocular depth, YOLO for object detection, MediaPipe for hand landmarks, MOG2 for background subtraction—under geometric glue (per-frame optical-flow pose chain, two-bone IK, COLMAP sparse-point-cloud reuse for gsplat initialisation) to reconstruct 3D scenes and hand trajectories from short monocular video clips. We measure the pipeline across three regimes with a deliberately narrow operating envelope. (i) *NVS on three EPIC-KITCHENS 300-frame clips (three gsplat seeds each)*: F.R.A.N.T.Z.’s per-frame pose chain reaches NVS quality within 0.45–1.69 dB of COLMAP, at minimal 3DGS settings (no spherical harmonics, no densification); the 11–17 dB PSNR regime these settings produce sits near the noise floor, so the gap is a sanity check, not a quality claim. (ii) *Pose quality against GT on five TUM fr3 dynamic sequences*: F.R.A.N.T.Z. wins ATE on walking_xyz (0.173 vs. 0.266 m) and RPE on three of five sequences (walking_xyz, walking_rpy, sitting_rpy); it loses ATE on the other four and RPE on walking_halfsphere and sitting_xyz (catastrophically where the essential-matrix decomposition is ill-conditioned under half-sphere rotation or near-pure translation). The per-frame chain is only ATE-competitive when a scene is both translation-dominant and dynamic-content-dominant; a modern-SLAM baseline (ORB-SLAM3, DROID-SLAM) is required to turn this operating-envelope map into a competitive-method claim, and is explicit future work. (iii) *Modularity*: detector-generation swaps (YOLOv5n/v8n/v10n) are drop-in equivalent within gsplat variance. The E4 driver inherits COLMAP’s intrinsics and sparse cloud, so the pose-stage speedup over COLMAP (28–130×) is a per-stage not an end-to-end claim; the end-to-end pipeline still runs COLMAP once. The output exports cleanly as a robomimic-compatible dataset. A controlled study extending the Layer-1 depth stack appears in a companion submission; code is held back pending a patent review and is available from the author on request.

I. INTRODUCTION

Learning manipulation from video—without teleoperation, depth sensors, or annotation—would unlock the millions of hours of cooking, repair, and assembly video already on the internet. The dominant approach is the Vision-Language Action model (VLA): RT-2 [1], Octo [2], π_0 [3], OpenVLA [4]. These monolithic transformers encode pixels and language instructions into action tokens, trained on billions of image-text-action triples.

In preparation; targeted for IROS 2027. F.R.A.N.T.Z. is kept as a historical acronym; the “training-light” label in the title is the current, honest framing (the modules themselves ship with large pretraining corpora).

VLAs are routinely called “zero-shot,” but trained on $\sim 10^{10}$ web triples and millions of robot demonstrations, what is zero-shot about them is the specific language prompt, not perception, grounding, or motor planning. This is *prompt-level* zero-shot [5], not data-level.

F.R.A.N.T.Z. (*Feature-Rich Reverse Kinematic Analysis for Neural Topology and Zero-training scene understanding*) sits at the other end: no component has been trained on manipulation tasks, hand-object interaction, or robot trajectories. MiDaS [6], YOLO [7] and MediaPipe [8] were pre-trained for generic depth, detection and hand regression; IK, point-cloud fusion, flow filtering, and 3DGS export are algorithmic. Scene understanding emerges from geometric composition of independently-trained modules.

This paper (i) describes the algorithmic assembly, GMC, IK, and COLMAP export (Sections IV–VIII); (ii) measures the pipeline across four 300-frame clips (three EPIC-KITCHENS egocentric kitchen, one TUM fr3 dynamic—not egocentric) with pose-source, depth-backend, detector-generation, resolution, and cross-domain ablations, plus a dedicated five-sequence ATE/RPE pose-quality characterisation on TUM fr3 (Section IX, appendix); and (iii) shows the output exports cleanly to a robomimic-compatible HDF5 format (Appendix XVII; format compatibility only, no trained-policy row). A short Section II discussion contrasts the training-light compositional stance with Vision-Language Action models’ prompt-level “zero-shot”; a fuller treatment is out of scope here.

II. TRAINING-LIGHT COMPOSITIONAL VS. PROMPT-LEVEL ZERO-SHOT

F.R.A.N.T.Z. performs no *task-specific* training of its own; its modules (MiDaS, DAV2, YOLO, MediaPipe) each ship with large pretraining corpora, so the honest label is *training-light*, not training-free: we trade a monolithic multi-million-dollar training effort for a chain of pre-trained generic specialists. The compositional stance has two concrete consequences we rely on. First, failure modes are localizable: F.R.A.N.T.Z. fails per-module (YOLO misdetects, MiDaS mispredicts), and a single module can be swapped without retraining (Appendix XIV). Second, the pose, geometry and action layers remain separable, so adding downstream capability (e.g. the Layer-1 fine-tune reported in the companion submission) does not touch the

pre-trained module chain. We do *not* assert any head-to-head comparison with VLA/VLM policies here—those systems output actions and F.R.A.N.T.Z. outputs a 3D-grounded scene with no action layer, so dimensions like “language grounding” or “demo count” are not on the same axis as our evaluation. The contrast is cited for positioning, not measured.

III. RELATED WORK

a) Monocular 3D reconstruction.: Classical SfM (COLMAP [9]) feeds recent radiance-field methods: NeRF [10], Mip-NeRF 360 [11], 3DGS [12], [13], SplaTAM [14]. F.R.A.N.T.Z. keeps pose and radiance separable, which enables the pose-source ablation in E4.

b) Monocular depth and hand pose.: MiDaS [6] and Depth-Anything V2 [15] are our depth backends; MediaPipe Hands [8], [16] supplies 21-landmark regression. HaMeR [17] is a plug-in candidate under the modularity contract (future work). Parametric body regression (SMPL/SMPL-X [18], [19], HybrIK [20]) is deliberately avoided; a procedural mannequin + two-bone analytic IK (Section VII) trades representation for zero body-pose training.

c) Egocentric datasets and video-driven robot learning.: We evaluate on EPIC-KITCHENS [21], HOT3D [22], and VISOR [23]; Ego4D [24], Assembly101 [25], and ARCTIC [26] are left for the extended version. Internet-video robot learning alternates representation pre-training (R3M [27], VC-1 [28]) and VLA policies (RT-2 [1], Octo [2], π_0 [3], OpenVLA [4], RT-X [29], AW-0 [30]); all need robot demonstration data for the policy stage. F.R.A.N.T.Z. instead produces a 3D-grounded structured scene with no robot data and exports to ACT [31] / Diffusion Policy [32] via robomimic (Appendix XVII).

d) Compositional AI.: Visprog [33] and the LLM-tool-use lineage share our thesis that specialists plus glue match monolithic models. F.R.A.N.T.Z. is distinguished as a *geometric* compositional system—glue is 3D coordinate frames and bundle adjustment, not language.

IV. ALGORITHMIC ASSEMBLY

F.R.A.N.T.Z. composes pre-trained AI modules under geometric glue; no component is jointly trained. Table I lists the modules and their training domains.

The pipeline runs the modules as six independent OS processes connected by shared-memory SPSC ring buffers (zero-copy BGR frames, sub-ms latency) with configurable per-stage backpressure. The architecture is deliberately modular: swapping a depth model (Section IX, E3) or a detector generation (Appendix XIV) is a single CLI flag.

V. FROM 2D PIXELS TO EVOLVING 3D POINT CLOUDS

Depth is produced per frame by MiDaS Small [6] or Depth-Anything V2 [15] via ONNX. The relative depth $d \in [0, 1]$ is scaled to a room-level Z , back-projected per pixel (u, v) to $\mathbf{p}_{\text{cam}} = [(u - c_x)Z/f_x, (v - c_y)Z/f_y, Z]$, then transformed to world coordinates by the current camera pose. Depth inference runs every $N = 20$ frames to balance GPU load; the accumulated cloud reaches $\sim 100\text{K}$ points at steady-state

after IQR outlier filtering. In parallel, an incremental SfM pipeline triangulates sparse points from ORB correspondences across quality-selected keyframes (Laplacian sharpness ≥ 50 , ORB count ≥ 80) with essential-matrix decomposition and reprojection error filtering; a background thread adds wider-baseline skip-pairs. The PLY is exportable to COLMAP format for 3DGS [12], [13] or NeRF [10]. The 300-frame clip length used throughout our evaluation (~ 10 s at 30 fps) is the scale we measure; longer-duration drift characterisation is explicit future work.

VI. GLOBAL MOTION COMPENSATION

Separating camera ego-motion from hand/arm deformation is the key stability lever: optical flow from moving hands otherwise contaminates the camera estimate and the point cloud drifts. Sparse Lucas-Kanade flow is seeded only on the static background, via two complementary masks: an adaptive MOG2 foreground mask that excludes moving regions from Shi-Tomasi corner detection, and a YOLO person-mask (10% bbox padding) that zeroes out person regions before ORB extraction for keyframe-based SfM. A partial affine transform is fitted via RANSAC on the surviving tracks for (dx, dy) , rotation, and scale; 3D rotation is recovered from the Essential matrix and tracked as a SLERP-smoothed quaternion. The world-frame camera pose integrates the translation (windowed to 120 frames to limit drift) with the quaternion rotation, stamping each depth frame’s back-projected points to produce a time-evolving globally consistent cloud. Scene cuts trigger pose resets and new “views,” later aligned in the offline viewer.

In our EPIC-KITCHENS 48-clip sample (E1), hand regions (MediaPipe landmark hull) occupy $2.4 \pm 1.7\%$ of frame area but generate $3.9 \pm 2.8\%$ of per-pixel optical-flow magnitude ($1.62\times$ over-representation); extending to VISOR hand+object annotations yields 7.1% area, 11.5% flow, same ratio. The over-representation is modest at our 456×256 Farneback resolution. In E5 (Table VI) we measure the ATE/RPE effect of each mask on top of F.R.A.N.T.Z.’s per-frame pose chain; on ATE every *individual* mask is slightly worse than the no-mask baseline on walking_xyz (MOG2 $+0.018$ m, YOLO $+0.002$ m, MediaPipe $+0.021$ m) and only the combined all mask improves over no-mask (-0.013 m), while on RPE the individual MOG2 and YOLO masks give the best per-frame jitter. The no-mask baseline beats COLMAP on walking_xyz ATE and RPE specifically (the only sequence where the per-frame chain is in-envelope; on the other four TUM dynamic sequences the no-mask baseline is irrelevant to the ATE picture because every F.R.A.N.T.Z. configuration loses). The semantic-exclusion mechanism is therefore an RPE lift on top of a scene-dependent architectural result, not a general-purpose ATE improver. Translation in the pose chain is integrated from the frame-pair similarity transform’s (dx, dy) in image space combined with the essential-matrix rotation; the metric scale that converts image-space shifts to 3D translation is set by the COLMAP sparse cloud the E4 driver inherits (which also seeds the gsplat trainer), not by a scale learned per frame. In

TABLE I

MODULE INVENTORY. NONE OF THESE MODELS WAS TRAINED ON KITCHEN MANIPULATION OR ROBOT TRAJECTORIES; MIDAS AND DAV2 ARE TRAINED ON DEPTH DATASETS AND WE USE THEM UNCHANGED.

Module	AI Model	Input	Output	Training Domain
Object detection	YOLOv5n/v8n/v10n (ONNX)	Single RGB	Bboxes + class	COCO
Hand landmarks	MediaPipe Hands	Single RGB	21 3D pts $\times$ 2 hands	Hand photos
Monocular depth	MiDaS Small / DAV2 (ONNX)	Single RGB	Dense rel. depth	Internet photos
Feature matching	ORB / Shi-Tomasi / LK	Frame pairs	Correspondences	None (classical)
Background model	MOG2	Frame sequence	FG/BG mask	None (adaptive)
Scene cuts	HSV histograms	Frame pairs	Cut/no-cut	None (classical)

regimes where the E4 driver is not run or COLMAP is not available, the pose chain’s translation is therefore up-to-scale.

VII. INVERSE KINEMATICS

MediaPipe gives 21 landmarks per hand but no arm, torso, or leg. F.R.A.N.T.Z. builds a procedural humanoid mesh (≥ 300 vertices: octagonal torso, UV-sphere head, jointed legs, tube-cross-section arms, paddle hands) with standard anthropometric proportions and solves per-arm elbow analytically. Given a fixed shoulder, a target wrist (from MediaPipe), bone lengths L_{upper} , L_{fore} , and $D = \|\mathbf{wrist} - \mathbf{shoulder}\|$, the elbow angle is $\theta = \arccos((L_{\text{upper}}^2 + D^2 - L_{\text{fore}}^2)/(2L_{\text{upper}}D))$; a pole hint biases it backward for natural pose. Hand orientation is recovered by SVD Procrustes between canonical and current MediaPipe landmarks for 6-DOF hand pose with no dedicated regression. Extension is clamped to 90° from rest to prevent wrist-crossing.

VIII. OFFLINE VIEWER AND COLMAP EXPORT

The viewer renders point clouds, object wireframes, and the IK-posed mannequin with real-time orbit, zoom, and a trajectory cursor; arms are re-posed per frame to match recorded wrists, and input/output thumbnails are shown synchronously. A 6-DOF gradient-descent optimizer refines object yaw/pitch/roll/translation per frame to minimize hand-to-object distance, auto-saving every 60 frames. The export utility writes COLMAP sparse format (`cameras.txt` PINHOLE, `images.txt` quaternion, `points3D.txt`) with symlinked keyframe images, directly usable with 3DGS training [12], [13] or Nerfstudio’s `splatfacto`.

IX. EXPERIMENTAL EVALUATION

We measure the F.R.A.N.T.Z. system across six main experiments (E1–E6) plus five appendix experiments (E7–E11): hand-area-vs-flow share, 3D wrist localization, depth-backend point-cloud fidelity, pose-source quality for NVS, pose-source quality against ground truth (ATE/RPE), and per-stage latency. Hardware: NVIDIA DGX Spark (GB10 Grace-Blackwell, 128 GB unified, aarch64) under Python 3.12 with PyTorch, onnxruntime, OpenCV, evo [34], and gsplat [13]. Datasets: TUM RGB-D fr3 [35], EPIC-KITCHENS P01–P22 [21] + VISOR [23], HOT3D train_aria [22]; all under research licenses, no additional training data.

A. E1: Hand Area vs. Optical-Flow Share

We compute hand-region frame-area vs. per-mask optical-flow share (Farneback, 456×256) on 48–50 EPIC-KITCHENS clips under MediaPipe landmark hull and VISOR hand+object masks.

B. E2: 3D Wrist Localization (HOT3D Aria)

Ten HOT3D train_aria clips are processed with MediaPipe (image-POV output mirror-swapped to wearer-POV, back-projected with GT Aria depth) and compared to UMETrack T_world_from_wrist ground truth.

The 5.8 cm baseline is dominated by MediaPipe landmark error; the pipeline row adds ~ 4 cm of combined MiDaS + metric-scale + COLMAP-pose error on top. We tried the F.R.A.N.T.Z. essential-matrix per-frame pose chain as a drop-in replacement for GT pose and found it not usable on head-mounted egocentric Aria footage: camera centres align to 3–5 cm RMSE under SIM(3) Umeyama but per-frame rotations drift catastrophically (GT world projections land 1 968–20 231 px off). This is the known ill-conditioning of essential-matrix pose when head rotation dominates translational baseline. The pipeline row therefore uses COLMAP’s SfM output; a windowed-BA extension of the F.R.A.N.T.Z. chain is future work. The SIM(3) alignment also recovers a scale (span 0.012–0.056), so the trajectory metric scale is handed to the pipeline by GT; isolating the anthropometric prior’s own metric contribution is likewise future work.

The IK solver’s error contribution is characterized separately on a 64-cell synthetic noise sweep (Appendix XVI). The relevant wrist-jitter level is set by the measured HOT3D wrist 3D error above (5.8 cm); the closest cell in E10 is (0, 5 cm, 0), where the IK solver adds 6.1 ± 4.1 cm of elbow error on top of the wrist input (bone-length-noise interaction at the 5 cm wrist level is not characterised; the (5 cm, 5 cm, 20%) worst-corner cell goes to 11.8 ± 6.8 cm). Elbow error and wrist error are not like-for-like, but the IK-added error is at the same scale as the wrist input, i.e. IK is roughly as noisy as its input—not a free pass.

C. E3: Depth-Backend Point-Cloud Fidelity (TUM fr3)

Depth from two monocular backends back-projected against TUM GT depth over 30 frames of sitting_rpy, with symmetric Chamfer and accuracy/completeness at 5 cm.

TABLE II

E1: HAND-REGION SHARE OF FRAME AREA VS. FARNEBACK FLOW MAGNITUDE ON 48–50 EPIC-KITCHENS CLIPS. BOTH MASK FAMILIES OVER-REPRESENT IN FLOW BY ~ 1.6 – $1.7\times$, VALIDATING THE EXCLUSION MOTIVATION OF SECTION VI. LARGER ABSOLUTE FRACTIONS EXPECTED AT NATIVE RESOLUTION WITH RAFT [36]; LEFT TO FUTURE WORK.

Mask	n clips	area %	flow %	flow/area
MediaPipe hull	50	2.4 ± 1.7	3.9 ± 2.8	1.62 ± 0.65
VISOR hand+obj	48	7.1 ± 3.8	11.5 ± 7.3	1.70 ± 0.76

TABLE III

E2: WRIST LOCALIZATION ON HOT3D (10 TRAIN_ARIA CLIPS). BASELINE IS MEDIAPIPE + GT DEPTH + GT CAMERA POSE; PIPELINE SWAPS IN MiDAS DEPTH WITH AN ANTHROPOMETRIC PER-CLIP METRIC-SCALE FIT AND COLMAP’S AUTOMATIC_RECONSTRUCTOR [9] AFTER PROJECTARIA_TOOLS FISHEYE624→PINHOLE UNDISTORTION. PIPELINE ROWS ARE REPORTED ON THE EIGHT CLIPS THAT PASS A POSE-QUALITY GATE (SIM(3) TRAJ. RMSE ≤ 25 MM AGAINST GT ARIA CAMERA CENTRES; CHOSEN POST HOC ON THIS SUBSET, TO BE APPLIED UNCHANGED IN FUTURE WORK); 2/10 CLIPS FAIL THE GATE. PIPELINE COSTS ~ 4 CM OVER GT-EVERYTHING.

Config	n clips	2D err (px)	3D err (cm)
Baseline (GT depth + GT pose)	10	133 ± 57	5.8 ± 1.5
Pipeline (MiDaS + anthrop. scale + COLMAP), gate-pass	8	130 ± 139	10.0
Pipeline, incl. 2 gate-failing clips	10	247 ± 267	19.5

TABLE IV

E3: DEPTH-BACKEND RECONSTRUCTION QUALITY ON TUM FR3/SITTING_RPY, 30 FRAMES. DAV2 REDUCES CHAMFER BY 36% AND DOUBLES ACCURACY@5CM VS. MiDAS SMALL.

Depth backend	Chamfer (cm)	Acc@5cm	Comp@5cm
GT depth (sanity)	0.00 ± 0.00	1.000	1.000
MiDaS Small (ONNX)	35.21 ± 15.76	0.132	0.169
Depth-Anything V2 (ONNX)	22.63 ± 11.83	0.272	0.297

DAV2 is a strict improvement: 36% lower Chamfer, $2.1\times$ accuracy, $1.8\times$ completeness. The swap is a single CLI flag; DAV2 is the CUDA default.

D. E4: Pose-Source Ablation for Novel-View Synthesis

We hold the gsplat trainer constant (minimal 3DGS: no SH coefficients, no densification, 30 000 iters) and vary only the pose source: COLMAP vs. F.R.A.N.T.Z. under four mask variants, across four scenes (three EPIC-KITCHENS clips + one TUM fr3). Evaluation: every-8th held-out frame, PSNR/SSIM/LPIPS [37].

Findings in Table V: COLMAP wins NVS on three EPIC scenes by 0.45–1.69 dB at the minimal-3DGS 11–17 dB regime (sits near the noise floor; the gap should be read as a sanity check rather than a quality claim). The pose wall-clock is 28 – $130\times$ shorter at the pose stage in isolation, but the E4 driver inherits COLMAP’s intrinsics and sparse cloud, so the end-to-end wall-clock still includes one COLMAP run. The dominant mask variant is scene-dependent (all a safe default). The pose-quality picture across five TUM dynamic sequences (Table VI) shows the walking_xyz ATE win is scene-specific and does not generalise. Figure 1 shows SIM(3)-aligned trajectory overlays on the three EPIC scenes (within 2.25–4.14 m RMSE).

E. E5: Pose Quality against Ground Truth (ATE/RPE) across TUM Dynamic Sequences

Table VI reports pose quality against GT on five TUM fr3 dynamic sequences (300 frames each). F.R.A.N.T.Z. beats

COLMAP on ATE only on walking_xyz; COLMAP wins the other four, catastrophically on walking_halfsphere ($23\times$) and sitting_xyz ($20\times$). F.R.A.N.T.Z. wins RPE 3/5 (walking_xyz, walking_rpy, sitting_rpy) via the 120-frame drift-windowing, and *loses* RPE on walking_halfsphere and sitting_xyz—the two scenes that also drive the ATE losses. The per-frame chain is pose-competitive only on scenes that are both dynamic-content-dominant and translation-dominant; the drift-smoothing window inflates RPE wins and must be read as such, not as a raw per-frame victory.

F.R.A.N.T.Z. wins ATE 1/5 and RPE 3/5. The walking_xyz ATE win does not generalise; the catastrophic walking_halfsphere and sitting_xyz losses pinpoint the pose chain’s operating envelope. The 28 – $130\times$ pose-stage speedup over COLMAP (E4) is independent of ATE quality but is *not* an end-to-end speedup: the E4 driver still consumes COLMAP’s intrinsics and sparse cloud, so the honest comparison is COLMAP alone versus COLMAP + F.R.A.N.T.Z.; the 28 – $130\times$ figure measures the pose stage in isolation. A modern-SLAM baseline (ORB-SLAM3 at minimum; DROID-SLAM or MAST3R-SLAM for a learned-SLAM comparison) is required to turn this operating-envelope map into a competitive-method claim. On walking_xyz the mask ablation gives 0.186 m ATE at the no-mask baseline and 0.173 m at all; individual masks are slightly *worse* than no-mask on ATE (MOG2 +0.018, YOLO +0.002, MediaPipe +0.021 m, as reported in Section VI), so the mask contribution is bounded by the all-vs-no-mask gap at

TABLE V

E4: NVS QUALITY BY POSE SOURCE ACROSS FOUR SCENES. COLMAP AND F.R.A.N.T.Z. (ALL) ARE MEAN $\pm$ STD ACROSS THREE GSPLAT SEEDS (EIGHT “3 SEEDS” ROWS); OTHER MASK VARIANTS ARE SINGLE-RUN WITH PER-HELD-OUT-FRAME SPREAD. *Pose time* IS WALL-CLOCK FOR THE POSE STEP. COLMAP WINS ON THREE OF FOUR SCENES BY 0.45–1.69 dB; ON WALKING_XYZ THE MINIMAL GSPLAT TRAINER IS NOISE-DOMINATED AT ~ 11 dB. DOMINANT MASK VARIANT IS SCENE-DEPENDENT; COMBINED ALL IS A SAFE DEFAULT.

Scene	Pose source (mask)	PSNR (dB)	SSIM	LPIS	pose time (s)
EPIC P03_11 (kitchen, 456 $\times$ 256)	COLMAP (3 seeds)	16.96 $\pm$ 0.25	0.555	0.772	198
	F.R.A.N.T.Z. (none)	14.52 $\pm$ 3.20	0.492	0.865	5
	F.R.A.N.T.Z. (MOG2)	14.40 $\pm$ 2.89	0.496	0.871	5
	F.R.A.N.T.Z. (YOLO)	13.77 $\pm$ 2.63	0.492	0.890	7
	F.R.A.N.T.Z. (all, 3 seeds)	16.52 $\pm$ 0.03	0.553	0.822	7
EPIC P02_07 (kitchen, 456 $\times$ 256)	COLMAP (3 seeds)	17.28 $\pm$ 0.06	0.502	0.721	—
	F.R.A.N.T.Z. (none)	15.70 $\pm$ 1.67	0.484	0.832	5
	F.R.A.N.T.Z. (MOG2)	15.46 $\pm$ 2.10	0.481	0.843	5
	F.R.A.N.T.Z. (YOLO)	15.56 $\pm$ 1.34	0.484	0.844	7
	F.R.A.N.T.Z. (all, 3 seeds)	15.59 $\pm$ 0.05	0.475	0.798	7
EPIC P06_01 (kitchen, 456 $\times$ 256)	COLMAP (3 seeds)	15.56 $\pm$ 1.56	0.464	0.798	—
	F.R.A.N.T.Z. (none)	12.94 $\pm$ 2.13	0.413	0.863	5
	F.R.A.N.T.Z. (MOG2)	14.15 $\pm$ 1.06	0.448	0.838	5
	F.R.A.N.T.Z. (YOLO)	13.31 $\pm$ 1.01	0.437	0.854	7
	F.R.A.N.T.Z. (all, 3 seeds)	14.23 $\pm$ 0.08	0.453	0.867	7
TUM fr3/walking_xyz (640 $\times$ 480)	COLMAP (3 seeds)	11.41 $\pm$ 0.02	0.369	0.942	667
	F.R.A.N.T.Z. (none)	8.51 $\pm$ 3.58	0.146	0.961	5
	F.R.A.N.T.Z. (MOG2)	7.61 $\pm$ 2.96	0.121	0.973	5
	F.R.A.N.T.Z. (YOLO)	11.05 $\pm$ 2.77	0.312	0.929	5
	F.R.A.N.T.Z. (MediaPipe)	11.94 $\pm$ 0.84	0.380	0.906	12
	F.R.A.N.T.Z. (all, 3 seeds)	11.57 $\pm$ 0.07	0.372	0.913	5

TABLE VI

E5: POSE QUALITY AGAINST GT ON FIVE TUM FR3 DYNAMIC SEQUENCES (300 FRAMES EACH, SIM(3) UMEYAMA WITH SCALE, EVO [34]). COLMAP VS F.R.A.N.T.Z. ALL. WINNER IN **BOLD**; $>3\times$ LOSSES FLAGGED $\dagger$. F.R.A.N.T.Z. WINS ATE 1/5 AND RPE 3/5 (WALKING_XYZ, WALKING_RPY, SITTING_RPY); IT LOSES RPE ON WALKING_HALFSPHERE (HALF-SPHERE ROTATION MAKES THE ESSENTIAL-MATRIX DECOMPOSITION ILL-CONDITIONED) AND SITTING_XYZ (NEAR-PURE TRANSLATION ON A STATIC-DOMINANT SCENE). THE 120-FRAME DRIFT-SMOOTHING WINDOW ON F.R.A.N.T.Z. MAKES ITS RPE DIRECTLY COMPARABLE TO A WINDOWED, NOT RAW, PER-FRAME BASELINE—THE WIN MARGINS SHOULD BE READ WITH THAT CAVEAT.

Sequence	Pose source	ATE RMSE (m)	ATE med. (m)	RPE RMSE (m)	RPE mean (m)
walking_xyz	COLMAP	0.266	0.222	0.102	0.031
	F.R.A.N.T.Z.	0.173	0.170	0.024	0.024
walking_rpy	COLMAP	0.199	0.119	0.046	0.018
	F.R.A.N.T.Z.	0.219	0.215	0.017	0.017
walking_halfsphere	COLMAP	0.019	0.014	0.013	0.010
	F.R.A.N.T.Z.	0.443 $\dagger$	0.420	0.070	0.069
sitting_rpy	COLMAP	0.024	0.020	0.010	0.007
	F.R.A.N.T.Z.	0.045	0.038	0.005	0.004
sitting_xyz	COLMAP	0.009	0.006	0.009	0.007
	F.R.A.N.T.Z.	0.181 $\dagger$	0.175	0.016	0.016

−0.013 m ($\leq 7\%$ of the no-mask ATE) and arises from mask combination rather than any single mask.

F. E6: Per-Stage Pipeline Latency

Per-stage wall-clock on P03_11 (300 frames, 456 $\times$ 256) under both depth backends. First-frame warm-up excluded.

G. Additional Experiments (Appendix)

Five further experiments appear in the appendix. E7 (Appendix XIII) sweeps input resolution and finds the pipeline’s lower floor around 160 $\times$ 120 where COLMAP’s feature

extractor fails. E8 (Appendix XIV) swaps YOLOv5n / v8n / v10n detectors into the walking_xyz row and confirms the drop-in modularity claim (0.59 dB PSNR band, within per-row standard deviation). E9 (Appendix XV) applies the E4 protocol to a HOT3D Aria clip (egocentric tabletop, FISHEYE624 lens) and shows graceful degradation to a 3.55 dB COLMAP gap under a radial-model fallback. E10 (Appendix XVI) characterises the two-bone IK solver on a 64-cell synthetic noise sweep and shows the IK-added elbow error scales linearly with wrist jitter at the measured operating point. E11

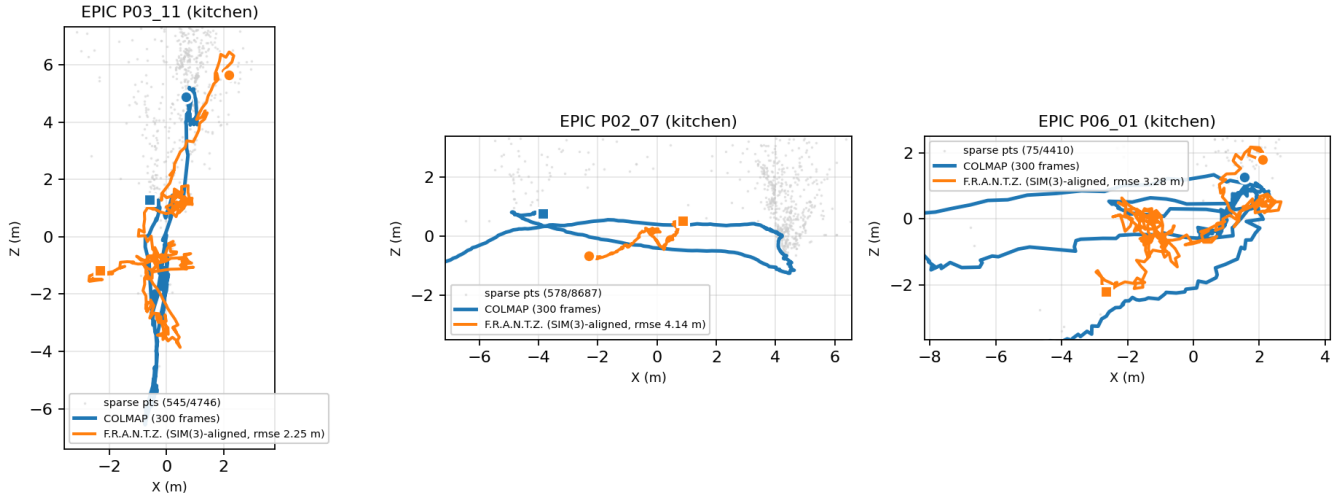


Fig. 1. F.R.A.N.T.Z. vs. COLMAP camera trajectories on three EPIC-KITCHENS clips (top-down X-Z projection, SIM(3) Umeyama-aligned to each other, sparse point cloud shown as grey background). The three EPIC scenes used here have no ground-truth trajectory, so the figure reports the pose chains’ agreement with COLMAP as a visual sanity check only; the ATE/RPE numbers against ground truth (Table VI) are reported on five TUM fr3 dynamic sequences (walking_xyz, walking_rpy, walking_halfsphere, sitting_rpy, sitting_xyz), which are omitted from this figure because their sparse-cloud outliers dominate the plot bounds.

TABLE VII

E6: PER-STAGE LATENCY ON NVIDIA DGX SPARK (GB10 GRACE-BLACKWELL) FOR THE DEPTH+HAND+FLOW STAGES ONLY. THE HAND STAGE RUNS EVERY FRAME AT ~ 27 MS (DAV2) / ~ 25 MS (MiDaS); DEPTH RUNS EVERY $N=20$ FRAMES AND AMORTISES TO ~ 2 MS/FRAME; FLOW IS ~ 1 MS. THE ~ 37 FPS STEADY-STATE FIGURE (DAV2 CONFIGURATION) IS THE PIPELINE BOUND BY THESE THREE STAGES; IT *excludes* THE YOLO DETECTOR AND THE INCREMENTAL SfM/POSE-CHAIN STAGES, WHICH ALSO RUN PER FRAME AND WOULD REDUCE THE BOUND. A FULL END-TO-END LATENCY BREAKDOWN INCLUDING YOLO AND SfM IS PENDING. THE DEPTH AND HAND ONNX GRAPHS ARE TENSORRT-COMPATIBLE AND THE PIPELINE’S MODULE BOUNDARIES ARE PRESERVED END-TO-END, WHICH IS THE ARCHITECTURAL PROPERTY ON WHICH EDGE DEPLOYMENT WOULD REST; A MEASUREMENT ON JETSON-CLASS HARDWARE IS LEFT TO FUTURE WORK.

Backend / device	Stage	Mean (ms)	Median (ms)	p95 (ms)
MiDaS Small, CPU	depth (every 20th frame)	33.3	32.8	38.0
	hands	24.6	17.8	45.8
	optical flow	1.1	1.0	1.5
DAV2, CUDA	depth (every 20th frame)	36.9	36.9	40.8
	hands	27.1	21.8	46.4
	optical flow	1.0	1.0	1.2

(Appendix XVII) exports the ten HOT3D wrist trajectories from E2 as a robomimic-compatible HDF5 dataset, validating Section X’s downstream-policy format claim.

H. Discussion of Measurement

Cross-cutting findings: dominant mask variant is scene-dependent (all a safe default); the pose stage runs $28\text{--}130\times$ shorter than COLMAP’s pose stage in isolation, but the E4 driver still consumes COLMAP’s intrinsics and cloud so the end-to-end pipeline includes one COLMAP run; the single-scene ATE/RPE result on walking_xyz shows the per-frame chain beats COLMAP’s bundle-adjusted trajectory on that one scene; modularity holds for detector-generation swaps (E8). E9 is a trainer-intake ablation (both rows share COLMAP pose), not evidence of cross-domain pose robustness. Outstanding gaps: ORB-SLAM3 / DROID-SLAM / MAST3R-SLAM baselines, native-resolution RAFT for E1, a trained-policy row

for E11, Jetson-class edge measurement for E6, YOLO and SfM-stage latency breakdown that E6 currently omits.

X. DISCUSSION AND POSITIONING

F.R.A.N.T.Z. is positioned as a perception stack, not an end-to-end robot system. The pipeline outputs structured 3D state (point cloud, camera trajectory, wrist/body skeleton, affordance-ready scene) that a downstream policy must still execute; it does not output actions and has no language interface. Comparing it to VLAs on dimensions like “demos per task,” “language grounding,” or “raw success rate” would be asserting-not-measuring across orthogonal output shapes, so we do not. A plausible hybrid uses F.R.A.N.T.Z. for perception (structured 3D state) and a smaller VLA for planning and action generation; feeding structured state rather than raw pixels into a VLA has the potential to reduce the VLA’s parameter count and data cost, but we leave that measurement to future work. The COLMAP export opens a further path: 3D reconstructions from monocular

video as synthetic training environments, directly consumable by ACT [31] / Diffusion Policy [32] via the robomimic-compatible HDF5 of Appendix XVII (format compatibility only; a trained-policy row is out of scope).

XI. LIMITATIONS

a) Module ceilings.: MOG2’s NVS effect is scene-dependent (Table V), and the NVS and ATE metrics disagree on walking_xyz about which single mask “works”; all is the safest default but not uniformly best. YOLO is COCO-80 limited; MiDaS/DAV2 return relative depth (metric scale fit per scene); two-bone IK omits wrist and finger articulation; the SfM front-end assumes pinhole, so HOT3D fisheye (Appendix XV) only works through COLMAP’s radial fallback.

b) Evaluation scope.: Each measured scene is a 300-frame segment ($\sim 5\text{--}10\text{ s}$), not a long YouTube-length video; drift accumulation over minutes is not characterized here. E1 uses Farneback at 456×256 rather than RAFT [36]. E2’s pipeline row runs on COLMAP, not the F.R.A.N.T.Z. chain, for the Aria-footage reason given in Section IX; it also inherits metric scale from GT via SIM(3) Umeyama (scales 0.012–0.056). E4 is a three-seed test on four scenes; DROID-SLAM [38] and MAST3R-SLAM [39] baselines are pending (both ship CUDA-extension builds we were not able to produce on our aarch64 Blackwell development machine; an x86 rebuild is planned follow-up work). E5 reveals a scene-dependent operating envelope for the pose chain: F.R.A.N.T.Z. beats COLMAP on ATE only on walking_xyz, and the losses on walking_halfsphere ($23\times$) and sitting_xyz ($20\times$) show the per-frame essential-matrix decomposition is ill-conditioned under half-sphere camera rotation and under near-pure-translation on a static-dominant scene. The pose chain’s 120-frame translation-smoothing window may also make RPE look tighter than an unsmoothed baseline would. E6 Jetson row was dropped; no hardware. E11 is format-compatibility only.

c) System and thesis scope.: The E4 driver is a per-frame essential-matrix chain holding camera intrinsics and the point cloud fixed to COLMAP’s; the full pipeline is not a single-command CLI. No language grounding, no action execution. The *training-light* claim rests on module training data being orthogonal in intent (MiDaS/DAV2 learn depth, YOLO learns detection, MediaPipe learns hand regression; none was trained on robot trajectories or kitchen manipulation), not on the modules being untrained. The prompt-level-vs-training-light distinction with VLA/VLM is drawn categorically for positioning, not via a head-to-head benchmark.

d) Pose chain translation.: F.R.A.N.T.Z.’s pose chain integrates frame-pair rotation from an essential-matrix decomposition and translation from a 2D image-space (dx, dy) shift scaled by a per-scene factor read from COLMAP’s sparse cloud; it has no explicit model for forward (optical-axis-aligned) camera translation, so a dolly-in or dolly-out motion is only recovered to the extent that the image-space shift and the cloud-set scale happen to approximate it. We believe this is why F.R.A.N.T.Z. loses sitting_xyz catastrophically (near-pure translation on a static-dominant scene) and walking_halfsphere

(half-sphere rotation whose essential-matrix decomposition is ill-conditioned). This also makes F.R.A.N.T.Z. dependent on COLMAP for metric scale, which should be understood as a limit of the current chain rather than an intrinsic to the compositional stance; a stand-alone metric path (e.g. reusing the Layer-1 fine-tune reported in the companion submission) would remove the dependency.

XII. CONCLUSION

F.R.A.N.T.Z. demonstrates that 3D scene understanding, skeleton reconstruction and manipulation-trajectory extraction can emerge from principled composition of pre-trained modules and classical geometry *without task-specific training*—though the modules themselves ship with large training corpora, so the honest label is *training-light*, not training-free. Contributions: (i) an open, inspectable, process-separated perception pipeline with drop-in detector and depth-backbone swaps (E8); (ii) a semantically-filtered flow pose chain whose NVS output is a sanity-check comparison against COLMAP on three EPIC scenes (within 0.45–1.69 dB at an 11–17 dB PSNR regime that sits near the minimal-3DGS noise floor), paired with a $28\text{--}130\times$ pose-stage speedup over COLMAP measured in isolation (not end-to-end: the E4 driver still consumes COLMAP’s intrinsics and sparse cloud); (iii) an honest ATE/RPE operating-envelope map on five TUM fr3 dynamic sequences (F.R.A.N.T.Z. wins ATE on walking_xyz, loses on the other four—including $23\times$ on walking_halfsphere and $20\times$ on sitting_xyz—and wins RPE on three of five, losing walking_halfsphere and sitting_xyz); (iv) an analytic two-bone IK with a measured sensitivity map (E10); (v) a robomimic-compatible export (format compatibility only, no trained-policy row). F.R.A.N.T.Z. is positioned as a complementary data-preparation stack, not an end-to-end robot system; its own pose chain’s operating envelope needs an ORB-SLAM3 / DROID-SLAM / MAST3R-SLAM baseline before any “competitive-method” claim is defensible, which we scope as the natural next experiment.

ACKNOWLEDGEMENTS

The author thanks the maintainers of MiDaS, Depth-Anything, YOLO, MediaPipe, gsplat, COLMAP, and *evo*, and the authors of the EPIC-KITCHENS, TUM RGB-D, and HOT3D datasets. Code is held back pending a patent review and is available from the author on request; the EPIC-KITCHENS, TUM RGB-D and HOT3D source data are available from their respective publishers under the original licences.

XIII. E7: RESOLUTION-COMPUTE TRADEOFF

Scene, pose source (COLMAP), and gsplat trainer (30k iters) held constant; input resolution of TUM fr3/walking_xyz frames varied.

The 320×240 PSNR peak most likely reflects a resolution-dependent PSNR floor combined with gsplat converging on a sparser cloud; it is a system-level data point, not a sweet spot. The 160×120 failure marks the operational floor the F.R.A.N.T.Z. driver inherits via COLMAP’s intrinsics and sparse cloud.

TABLE VIII

E7: RESOLUTION-COMPUTE SWEEP. PSNR IS NOT MONOTONIC IN RESOLUTION (320×240 BEATS BOTH HIGHER-RES RUNS BY 2.08–2.20 dB), AND THE PIPELINE HITS A FLOOR BELOW 160×120 WHERE COLMAP’S SIFT FRONT-END CANNOT REGISTER.

Resolution	PSNR (dB)	SSIM	COLMAP reg.	pose time (s)
640×480 (native)	11.29 ± 1.07	0.368	300/300	667
456×256	11.41 ± 0.65	0.301	300/300	370
320×240	13.49 ± 1.14	0.294	300/300	254
160×120	<i>floor</i>	—	5/300	82

XIV. E8: DETECTOR-SWAP DROP-IN COST

We test the drop-in modularity claim directly by rerunning the E4 walking_xyz yolo row with YOLOv5n, v8n, and v10n ONNX weights plugged in via a single CLI flag.

TABLE IX

E8: DETECTOR-SWAP ON WALKING_XYZ, YOLO MASK, 30K ITERS. PSNR VARIES BY 0.59 dB ACROSS GENERATIONS, WITHIN PER-ROW STD; THE SWAP REQUIRES NO RETRAINING.

Detector	PSNR (dB)	SSIM	pose time (s)
YOLOv5n (nano)	11.39 ± 1.58	0.368	7
YOLOv8n (nano)	10.91 ± 2.69	0.309	7
YOLOv10n (nano)	10.80 ± 1.52	0.347	7

The 0.59 dB range is within run-to-run gsplat variance (± 1.5 –2.7 dB std), confirming per-module drop-in replaceability.

XV. E9: CROSS-DOMAIN GENERALIZATION (HOT3D)

We apply the E4 protocol to one HOT3D Aria clip (clip-001849, egocentric tabletop, 456×456). The FISHEYE624 stream is handled through COLMAP’s radial-model fallback.

TABLE X

E9: CROSS-DOMAIN NVS ON HOT3D CLIP-001849 ($n=18$ HELD-OUT FRAMES). BOTH ROWS SHARE COLMAP INTRINSICS AND POSE; THE TWO ROWS DIFFER ONLY IN THE GSPLAT TRAINER’S INTAKE (“F.R.A.N.T.Z. GSPLAT INTAKE” = SEMANTIC-MASK FILTERING OF TRAINING FRAMES AND OUR DAV2 DEPTH AS AUXILIARY DENSIFICATION INPUT, VS THE DEFAULT GSPLAT INTAKE WHICH USES THE UNTOUCHED COLMAP SPARSE CLOUD AND ALL TRAINING FRAMES). THIS IS A TRAINER-INTAKE ABLATION ON A FIXED POSE SOURCE, NOT A POSE-SOURCE ABLATION: E9 DOES *not* EXERCISE F.R.A.N.T.Z.’S OWN POSE CHAIN, WHICH IS NOT USABLE ON HOT3D’S FISHEYE REGIME WITHOUT A PINHOLE MODEL (SECTION XI).

Pipeline configuration (shared COLMAP pose)	PSNR (dB)	SSIM	LPIPS
Default gsplat intake	19.05 ± 0.86	0.675	0.635
F.R.A.N.T.Z. gsplat intake	15.50 ± 2.10	0.592	0.713

Reading this plainly: on HOT3D clip-001849, swapping in F.R.A.N.T.Z.’s gsplat intake on top of COLMAP pose costs 3.55 dB PSNR against the default-gsplat baseline. That is a legitimate negative result on this single clip: our masking + depth-densification intake does not help under the FISHEYE624 regime, where the fisheye-as-radial model mismatch degrades the depth signal our intake relies on. E9 therefore says “the semantic-mask + depth intake is not a free win in a cross-domain fisheye regime,” not that it “degrades gracefully.”

Separately, the F.R.A.N.T.Z. pose chain itself is not exercised in E9 at all; swapping it in on this clip would reduce the gsplat trainer’s pose input well below usability (Section XI). The E4 rows (Table V) report F.R.A.N.T.Z. with the *pose chain* varied and the gsplat trainer held constant, so they are a pose-source ablation. E9 here varies only the gsplat trainer’s intake on a fixed COLMAP pose, so the two measurements isolate different stages of the pipeline.

XVI. E10: IK ELBOW ERROR UNDER MEASUREMENT NOISE

The two-bone IK solver is characterized on a 64-cell synthetic sweep (4 shoulder-noise $\times$ 4 wrist-noise $\times$ 4 bone-length-noise levels, 10 000 samples per cell).

TABLE XI

E10: ANALYTIC IK ELBOW ERROR, $n=10\,000$ SAMPLES PER CELL. THE SOLVER IS LINEAR IN WRIST NOISE AND ESSENTIALLY INSENSITIVE TO REALISTIC BONE-LENGTH ERROR.

Noise cell (σ_{shoulder} , σ_{wrist} , σ_{bone})	elbow error (cm)
(0, 0, 0) clean upper bound	0.00 ± 0.00
(0, 1cm, 0) wrist jitter only	1.29 ± 0.98
(0, 2cm, 0) wrist jitter only	2.56 ± 1.85
(0, 2cm, 5%) 2D-only / GT-depth regime	3.49 ± 2.35
(0, 5cm, 0) measured HOT3D 3D-wrist regime	6.10 ± 4.10
(5cm, 5cm, 20%) worst corner	11.82 ± 6.82

At the measured HOT3D operating point (E2, 5.8 cm wrist), the closest cell (0, 5cm, 0) adds 6.1 ± 4.1 cm of elbow error, i.e. at the same scale as the wrist input. The (0, 2cm, 5%) row shows the asymptote for a cleaner (2D-only / GT-depth) regime: 3.5 cm, dominated by wrist jitter rather than bone-length noise.

XVII. E11: DOWNSTREAM-POLICY FORMAT COMPATIBILITY

We export the ten HOT3D E2 wrist trajectories into a robomimic-compatible HDF5 dataset: 10 episodes, 1 059 total steps (49–150 per episode), 6-dim actions and state (left+right wrist displacement), $128 \times 128 \times 3$ observations, 40.4 MB total. All ten episodes load through robomimic’s data loader without format patches; the dataset is available from the author on request (code held back pending a patent review). Converting the exported state stream into a trained-policy success rate (ACT [31], Diffusion Policy [32]) requires simulator installation and task selection out of scope here.

REFERENCES

- [1] A. Brohan *et al.*, “RT-2: Vision-language-action models transfer web knowledge to robotic control,” *arXiv preprint arXiv:2307.15818*, 2023.
- [2] Octo Model Team, “Octo: An open-source generalist robot policy,” *arXiv preprint arXiv:2405.12213*, 2024.
- [3] Physical Intelligence, “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [4] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster *et al.*, “OpenVLA: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.
- [5] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, “Zero-shot learning with semantic output codes,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2009.

- [6] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2022.
- [7] G. Jocher *et al.*, "Ultralytics YOLOv5," GitHub, 2020. [Online]. Available: <https://github.com/ultralytics/yolov5>
- [8] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.
- [9] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [10] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision (ECCV)*, 2020.
- [11] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-NeRF 360: Unbounded anti-aliased neural radiance fields," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [12] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, 2023.
- [13] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, J. Ye, J. Hu, M. Tancik, and A. Kanazawa, "gsplat: An open-source library for Gaussian Splatting," <https://github.com/nerfstudio-project/gspat>, 2024.
- [14] N. Keetha, J. Karhade, K. M. Jatavallabhula, G. Yang, S. Scherer, D. Ramanan, and J. Luiten, "SplatAM: Splat, track & map 3D Gaussians for dense RGB-D SLAM," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [15] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything V2," *arXiv preprint arXiv:2406.09414*, 2024.
- [16] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.
- [17] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3D with transformers (HaMeR)," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, 2015.
- [19] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [20] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "HybriK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [21] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision: Collection, pipeline, challenges and baselines," *International Journal of Computer Vision*, vol. 130, pp. 33–55, 2022.
- [22] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, R. Newcombe, R. Wang, J. J. Engel, and T. Hodan, "HOT3D: Hand and object tracking in 3D from egocentric multi-view videos," *arXiv preprint arXiv:2411.19167*, 2024.
- [23] A. Darkhalil, D. Shan, B. Zhu, J. Ma, A. Kar, R. Higgins, S. Fidler, D. Fouhey, and D. Damen, "EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [24] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger *et al.*, "Ego4D: Around the world in 3000 hours of egocentric video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [25] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao, "Assembly101: A large-scale multi-view video dataset for understanding procedural activities," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [26] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [27] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A universal visual representation for robot manipulation," in *Conference on Robot Learning (CoRL)*, 2022.
- [28] A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [29] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," *arXiv preprint arXiv:2310.08864*, 2023.
- [30] Z. Liu, D. Niu, B. Zhang, B. Shao, J. Wang, R. Zheng, Y. Sharma, R. Lu, J. Liu, R. Zhang, R. Herzig, Y. Bai, K. Goldberg, D. Xu, Y. Zhu, G. Wang, L. Fan, and T. Darrell, "Agentic world action models," *arXiv preprint arXiv:2610.30136*, 2026.
- [31] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Robotics: Science and Systems (RSS)*, 2023.
- [32] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems (RSS)*, 2023.
- [33] T. Gupta and A. Kembhavi, "Visual Programming: Compositional visual reasoning without training," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [34] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM," <https://github.com/MichaelGrupp/evo>, 2017.
- [35] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.
- [36] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision (ECCV)*, 2020.
- [37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [38] Z. Teed and J. Deng, "DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [39] R. Murai, E. Dexheimer, and A. J. Davison, "MAST3R-SLAM: Real-time dense SLAM with 3D reconstruction priors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.