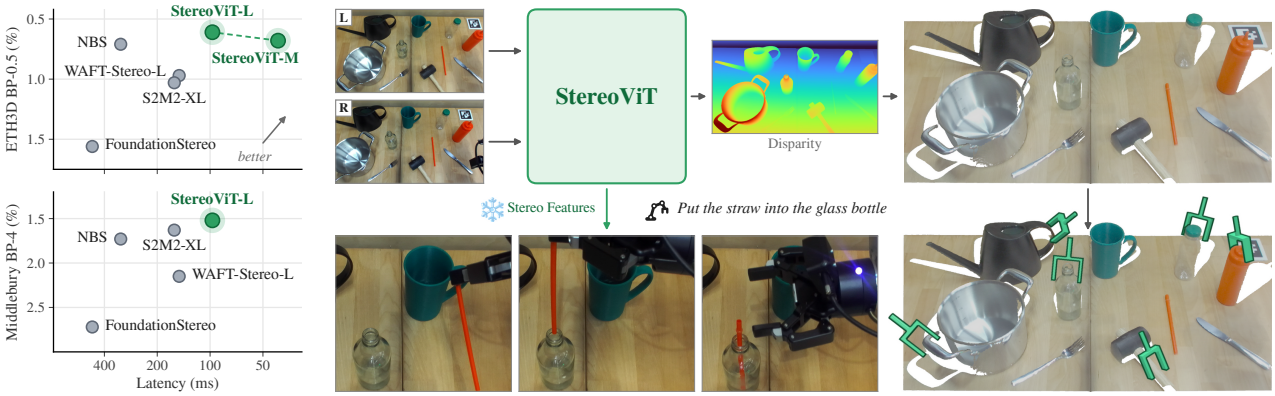
arXiv:submit/8431958 [cs.CV] 26 Nov 2026

StereoViT: Stereo Vision Transformer for Robust Robot Manipulation

Dingxi Zhang[1,2] Alexey Gavryushin[1] Bernt Schiele[2] Marc Pollefeys[1,3] Haofei Xu[1,4]
[1]ETH Zurich [2]Max Planck Institute for Informatics, Saarland Informatics Campus
[3]Microsoft [4]KE:SAI

https://kristen-z.github.io/projects/stereovit



(a) Disparity accuracy vs. latency (b) Frozen features for policy learning (c) Point cloud for grasping

Figure 1. **StereoViT is a streamlined stereo model that is accurate, fast, and transferable to robust robot manipulation.** (a) Disparity accuracy vs. latency on ETH3D and Middlebury, comparing StereoViT (M/L) with prior work. (b) From a stereo pair, StereoViT predicts disparity in one forward pass, providing frozen features for policy learning (e.g., inserting a straw into a bottle). (c) The predicted metric point cloud captures transparent, reflective, and thin objects, enabling direct grasping via EconomicGrasp [111].

Abstract

Stereo cameras give robots metric geometry from passive, cheap sensors, and stereo input has been shown to improve manipulation policies over monocular RGB. Yet unlike monocular vision, stereo lacks a widely adopted feature backbone for robot learning: leading stereo models rely on cost volumes, feature warping, and iterative refinement, so their geometry emerges over many refinement steps, leaving intermediate features less suited for direct reuse. Inspired by recent evidence that these inductive biases are not strictly necessary for accuracy, we present StereoViT, a streamlined stereo vision transformer that provides both accurate disparity and transferable stereo features for robotic manipulation. StereoViT first uses a depth-pretrained ViT to encode each view, and then a transformer decoder to relate the stereo pair with attention, predicting disparity in a single feed-forward pass. With one checkpoint, StereoViT ranks first on KITTI 2012 and 2015 (all metrics), ETH3D (BP-0.5), Middlebury (BP-4), and Spring (all metrics) and is 1.5–4.8× faster than competing methods. Its depth yields the highest grasp success among stereo methods, while its frozen stereo features match the strongest diffusion policy baselines in clean scenes and achieve much higher success rates under domain randomization on the RoboTwin 2.0 benchmark. Real-world experiments further confirm these gains in unseen scenes.

1. Introduction

Visuomotor policies, which map camera observations directly to robot actions, have become a leading paradigm for learning manipulation from demonstrations [2, 10, 11, 19, 51, 88, 122, 128]. Their progress relies on pretrained visual encoders, which turn raw pixels into features that support policy learning from limited demonstrations [66, 73, 85]. As manipulation becomes more precise and contact-rich, these features are increasingly incorporating geometric information, drawing on monocular depth estimation and multi-view reconstruction [74, 80, 92, 123, 127].

Stereo offers a promising yet underexplored source of geometric information for robot learning. A pair of cameras provides metric 3D structure from a single observation, without relying on active illumination that can fail on transparent and specular surfaces [20, 23, 82]. Stereo also operates directly on images, allowing pretrained visual en-

coders to be reused [36, 108]. Recent studies have shown that stereo observations can improve manipulation policies over monocular counterparts [21, 36, 38, 61].

Despite these advantages, stereo foundation models [8, 34, 108] are primarily developed for disparity estimation, with less emphasis on producing general-purpose features for downstream robot learning. Existing approaches therefore leverage stereo in different ways. StereoVLA [21] uses features from the cost volume of FoundationStereo [108], while StereoPolicy [36] finds that frozen FoundationStereo features underperform stereo fusion learned from task demonstrations. Meanwhile, ExStereo [61] favors explicit point clouds as input. As these approaches also differ in their policy architectures, it remains unclear how effectively pretrained stereo features and explicit geometry can support robot manipulation under a common policy framework.

Bridging this gap calls for a stereo backbone that provides accurate disparity while yielding transferable features that can be reused by a robot. However, leading stereo models are typically built around disparity-specific operations: cost volumes enumerate disparity hypotheses, feature warping aligns features according to current estimates, and iterative updates progressively refine the prediction [15, 49, 60, 70, 101, 113, 116]. Although effective for disparity estimation, these pipelines rely on successive refinement and task-specific engineering components, rather than directly producing reusable features. Iterative refinement also adds latency, an important consideration for robotics [21, 36, 38].

Recent work suggests that much of this machinery is not strictly necessary for accurate disparity estimation. NBS [93] and DispViT [31] demonstrate the effectiveness of stereo models built on plain vision transformers [22]. By removing engineered, task-specific components, such architectures offer a simpler and potentially more generalizable foundation for stereo vision. However, accuracy gaps remain relative to specialized stereo methods on benchmarks such as Middlebury [83], and whether their learned features can effectively transfer to robotics remains largely unexplored. Inspired by these advances, we present StereoViT, a streamlined stereo vision transformer that achieves leading disparity accuracy across five benchmarks using one checkpoint and provides transferable stereo features for robust robotic manipulation.

Specifically, StereoViT uses a shared ViT initialized from Depth Anything V2 [119] to encode each view, providing depth-aware features where stereo correspondence is ambiguous. A transformer decoder adapted from FreeFlow [7] relates the views through window, shifted-window [63], and global attention. In the global pass, cross-attention is restricted to epipolar rows on a pooled token grid, focusing interaction where correspondences lie. A lightweight head predicts disparity as an expectation over disparity anchors, while the decoder features are kept frozen for downstream policy learning.

With a single checkpoint, StereoViT ranks first on five major stereo benchmarks: KITTI 2012 [27], KITTI 2015 [69], ETH3D (BP-0.5) [84], Middlebury (BP-4) [83], and Spring [68]. StereoViT-L runs 1.5–4.8× faster than competing high-accuracy methods, while StereoViT-M is 11× faster than FoundationStereo. Beyond disparity estimation, its depth yields the highest grasp success among stereo methods on DREDS [20], particularly for transparent objects. Its frozen stereo features match the strongest baselines in clean RoboTwin 2.0 [16] scenes and achieve 1.7× the success rate of the next best method under domain randomization. Real-world experiments further demonstrate its applicability to robust manipulation.

In summary, our contributions are:

- **A streamlined stereo vision transformer.** StereoViT predicts disparity in a single pass, eliminating cost volumes, feature warping, and iterative refinement.
- **State-of-the-art accuracy and efficiency.** A single checkpoint tops five stereo benchmarks while running 1.5–4.8× faster than competing methods.
- **Transferable features for robotic manipulation.** StereoViT improves depth-based grasping while its frozen features enable robust policy learning under visual variations in simulation and real-world experiments.

2. Related Work

Stereo disparity estimation. Classical stereo aggregates correspondence costs under smoothness constraints [40]. Learning-based methods progress from direct disparity regression [67] to cost volumes filtered by 3D convolutions [15, 33, 49, 86, 115]. RAFT-Stereo [60] introduces recurrent updates over correlation volumes, a paradigm refined by subsequent methods [17, 29, 55, 101, 113, 114]. Recent approaches improve generalization by incorporating monocular depth priors [8, 18, 30, 45, 108] and scaling training data [34, 103]. Others explore attention-based cross-view interaction [32, 57, 70, 107, 116, 126] or feature warping [103], often retaining iterative refinement. More recently, NBS [93] and DispViT [31] demonstrate strong stereo estimation with streamlined ViTs, while GeoViT [110] explores adapting pretrained video transformers to multi-view geometry. FreeFlow [7] introduces a hierarchical transformer for optical flow. Building on these advances, StereoViT combines depth-pretrained per-view encoding, a cross-view decoder whose global cross-attention follows epipolar rows, and an expectation-based disparity head. Unlike prior streamlined stereo models evaluated primarily for disparity estimation, we further explore both explicit depth and frozen stereo features for robust robot manipulation.

Depth for robotic grasping. Stereo disparity yields met-

ric depth, which can be back-projected into point clouds for 6-DoF grasp detection [25, 26, 35, 72, 98, 111]. Grasping performance therefore depends on the quality of the estimated depth. Transparent and specular objects are particularly challenging, often causing missing or corrupted depth measurements from active sensors [20, 82]. Prior work addresses these failures through depth completion [23, 82, 87], monocular estimation [62, 121], or stereo-based depth estimation [4, 5, 64, 71, 106]. We evaluate StereoViT as a drop-in depth source on the material-diverse DREDS benchmark [20], keeping the grasp detectors fixed to isolate the effect of stereo depth quality.

Visual representations for manipulation. Complementing depth-based grasp detection, pretrained visual features enable robot policies to learn from limited demonstrations [66, 73, 85]. Vision-language-action models inherit semantic priors from vision-language models [10–12, 51, 88], while world-action models leverage temporal priors from video generation or predictive feature learning [2, 3, 52, 76, 122, 130, 131]. To incorporate geometric information, robot policies consume point clouds or depth-lifted image features [28, 43, 44, 47, 120, 124, 125], and vision-language-action models increasingly incorporate depth and 3D inputs [53, 56, 59, 80, 123, 129]. More recently, geometric foundation models [48, 58, 99, 105] have been adopted for robot learning through feature alignment [54, 91], encoder replacement [1, 74, 79, 127], point-cloud fusion [92], and joint geometry-action prediction [37, 90, 112]. These models primarily target monocular or general multi-view geometry. Stereo vision offers another source of geometric information, and recent work has demonstrated its benefits for manipulation. StereoPolicy [36] learns stereo fusion from task demonstrations, StereoVLA [21] incorporates pretrained features [108] into a VLA model, ExStereo [61] uses stereo-derived point clouds, and EyeRobot 2.0 [38] combines stereo with active gaze. While these approaches integrate stereo into different robotic pipelines, we investigate whether a stereo backbone trained for disparity estimation can provide useful features for visuomotor policies.

3. Method

Given a rectified stereo pair $I_l, I_r \in \mathbb{R}^{H\times W\times 3}$, StereoViT estimates the left-view disparity $d \in \mathbb{R}^{H\times W}$ in a single forward pass. Metric depth follows as $Z = fB/d$, where f is the horizontal focal length in pixels and B is the camera baseline. The model comprises a shared image encoder $\mathcal{E}$, a cross-view decoder $\mathcal{D}$, and a disparity head $\mathcal{H}$ (Fig. 2):

$$F = \mathcal{D}\big(\mathcal{E}(I_l), \mathcal{E}(I_r)\big), \qquad d = \mathcal{H}(F). \tag{1}$$

The decoder produces a C-channel stereo feature map $F \in \mathbb{R}^{\frac{H}{14}\times\frac{W}{14}\times C}$, which we use for downstream policy learning with the stereo backbone frozen.

3.1. Network architecture

The encoder aggregates multi-layer pretrained image features, the decoder integrates cross-view information through local and global attention, and the head predicts continuous disparity using fixed anchors.

Monocular encoder. Cross-view evidence can be ambiguous in textureless or occluded regions and on transparent or reflective surfaces, where monocular cues such as shape, layout, and context provide complementary information. We therefore encode each view independently with a shared ViT initialized from Depth Anything V2 (DAv2) [119] and fine-tuned during stereo training. The encoder uses its native 14×14 patch size. Following NBS [93], we upsample the pretrained 37×37 positional embeddings to 148×148 and fine-tune them. Features from four evenly spaced encoder layers are concatenated and linearly projected to the decoder width C.

Cross-view decoder. We adopt an asymmetric decoder following FreeFlow [7] and CroCo [107]. Only the left-view tokens are updated; the right-view tokens remain fixed and provide context for cross-attention. To combine fine spatial detail with long-range context efficiently, each of the N blocks applies window, shifted-window, and pooled global attention in sequence. Each pass consists of self-attention over the left-view tokens, cross-attention to the right-view tokens, and an MLP:

- **Window Attention.** The token grid is split into 4×4 non-overlapping windows, and attention is computed within each window, capturing local detail.
- **Shifted-window Attention.** The windows are shifted by half their size to exchange information across window boundaries [63].
- **Global Attention.** A stride-2 convolution downsamples both views by a factor of 2. Self-attention spans the entire left grid, while cross-attention is restricted to each token's row of the right view, the epipolar line on which its correspondence lies. A transposed convolution upsamples the result, which is added back to the tokens.

Disparity head. In cost-volume methods, the soft argmin over disparity hypotheses is a natural readout, as each hypothesis is a matching candidate [49, 108]; without a cost volume, direct regression [7, 107] could be another way. Instead, we read out the disparity as an expectation over fixed anchors, inspired by bin-based monocular depth estimation [9]. For each decoder token, a lightweight convolutional head predicts probabilities p_j over $M = 128$ fixed disparity anchors $c_j = 8(j-1)$ px, spanning 0 to 1016 px. The token-resolution disparity is $\widetilde{d} = \sum_{j=1}^{M} p_j c_j$. The output is thus bounded and continuous between anchors, and the network only has to place probability mass rather than produce a value at pixel scale. Convex upsampling [94] produces the full-resolution disparity map d. Compared with

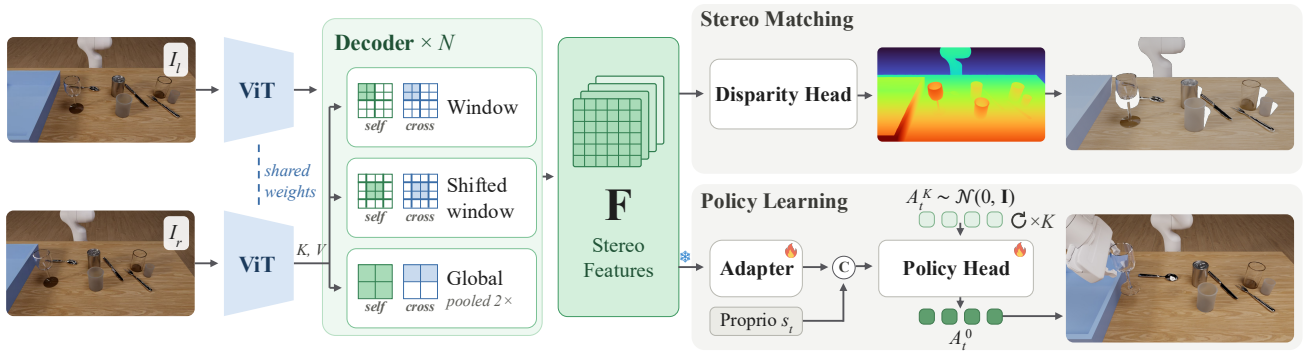


Figure 2. **Overview of StereoViT.** A shared ViT encoder, initialized from Depth Anything V2 [119], embeds each view. A transformer decoder relates the two views through window, shifted-window and global attention, producing the stereo feature F in a single forward pass. A light head reads disparity from F, and the same F conditions a diffusion policy for manipulation.

direct regression, this readout converges faster and reaches about half the training error at the same budget, which carries over to zero-shot accuracy (Tab. 6).

3.2. Training for disparity estimation

Model variants. We train two variants with the same recipe. StereoViT-M uses a ViT-B encoder with decoder depth $N = 6$ and width $C = 384$, while StereoViT-L uses a ViT-L encoder with $N = 8$ and $C = 512$.

Supervision. We supervise the final full-resolution disparity using the mixture-of-Laplace loss $\mathscr{L}_{\text{MoL}}$ [102–104]. The disparity head additionally predicts the mixture parameters required by this loss.

Training recipe. We train both variants on 4.5M synthetic stereo pairs from 12 public sources [13, 14, 55, 67, 68, 77, 96, 97, 100, 108, 117, 118]. We denote this mixture as SynLarge following WAFT-Stereo [103]. The resulting checkpoints are used for zero-shot stereo evaluation. For official benchmark submissions, we fine-tune StereoViT at higher resolution. Following [103], the fine-tuning mixture combines a subset of SynLarge with the training splits of Middlebury [83] and KITTI [27, 69], together with Booster [81] and InStereo2K [6]. The same fine-tuned checkpoint is submitted to KITTI 2012 and 2015, Middlebury, ETH3D [84], and Spring [68]. No ETH3D data are used for fine-tuning.

3.3. Transfer to robotic manipulation

We evaluate StereoViT in two downstream settings: grasp detection from predicted depth and visuomotor policy learning from stereo features. The stereo backbone is kept frozen in both settings.

Depth-based grasp detection. We convert the predicted disparity to metric depth and back-project it into a point cloud using the camera intrinsics. The resulting point cloud serves as input to off-the-shelf 6-DoF grasp detectors [98, 111]. For each detector, we keep its weights and the downstream grasping pipeline fixed while varying the depth source. This evaluates the utility of the predicted geometry without introducing a new grasping method. Depth and grasping evaluation protocols are described in Sec. 4.2.

Feature-based policy learning. At each time step t, the frozen StereoViT backbone extracts stereo features F_t from the input pair. We average-pool F_t to a fixed spatial grid and map it to a 512-dimensional observation $z_t = g(F_t)$ using a lightweight convolutional adapter g. The adapter has 1.1M parameters, an order of magnitude fewer than the ResNet-18 [39] visual encoder (11.2M) it replaces in Diffusion Policy (DP) [19]. We use the convolutional variant of DP to predict action chunks of length T_a, conditioned on the visual features and proprioceptive states from the last n observations:

$$O_t = \big[z_{t-n+1} \,\|\, s_{t-n+1} \,\|\, \cdots \,\|\, z_t \,\|\, s_t\big]. \tag{2}$$

The noise predictor ε_θ is a 1D temporal U-Net conditioned on O_t and the diffusion step k through FiLM layers [78]. During training, a ground-truth action chunk $A_t^0 = (a_t, \ldots, a_{t+T_a-1})$ is corrupted with Gaussian noise, and the policy learns to predict the added noise [41]:

$$\begin{aligned} A_t^k &= \sqrt{\bar{\alpha}_k}\, A_t^0 + \sqrt{1-\bar{\alpha}_k}\, \epsilon, \\ \mathcal{L}_{\text{DP}} &= \mathbb{E}_{k,\epsilon}\Big[\big\|\epsilon - \varepsilon_\theta(O_t, A_t^k, k)\big\|_2^2\Big], \end{aligned} \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $k \in \{1, \ldots, K\}$ is the sampled diffusion step, and $\bar{\alpha}_k$ is determined by the noise schedule. Only the adapter g and noise predictor ε_θ are trained. At inference, the policy starts from $A_t^K \sim \mathcal{N}(0, \mathbf{I})$ and denoises it over K steps to obtain an action chunk A_t^0. The robot executes an initial portion of the chunk before replanning. More details are provided in the supplementary.

4. Experiments

We evaluate StereoViT as a disparity model in Sec. 4.1, as a depth source for grasping in Sec. 4.2, as a frozen repre-

Table 1. **Official test benchmarks and efficiency.** Baseline metrics are sourced from official leaderboards. StereoViT-L uses a single fine-tuned checkpoint across ETH3D, KITTI, Middlebury and Spring. (†): zero-shot evaluation on ETH3D. (–): no entry. All pixels are evaluated; lower is better. Parameters, latency, and peak memory are tested on a 544×960 input using a single GH200 GPU in FP32.

Method	ETH3D			Middlebury			KITTI 2012		KITTI 2015		Spring		Params	Latency	Memory
	BP-0.5	BP-2	EPE	BP-2	BP-4	EPE	BP-2	BP-3	D1-all	D1-bg	1px	D1	(M)	(ms)	(GB)
RAFT-Stereo [60]	7.33	0.56	0.19	9.37	6.42	2.71	–	–	2.42	1.92	15.27	8.63	11	139	1.63
UniMatch [116]	6.44	0.34	0.21	11.7	6.07	1.89	–	–	1.77	1.49	–	–	7	53	1.56
CroCo-Stereo [107]	3.51	0.50	0.15	11.1	6.75	2.36	–	–	1.59	1.38	7.14	2.71	437	1100	2.14
Selective-IGEV [101]	3.46	0.51	0.15	6.04	3.74	1.54	2.05	1.38	1.55	1.33	–	–	13	199	1.00
BridgeDepth [30]	1.76	0.16	0.11	6.92	3.48	1.48	1.65	1.03	1.40	1.13	–	–	346	97	1.89
FoundationStereo [108]	1.56	0.26	0.13	4.26	2.72	1.24	–	–	–	–	–	–	375	469	4.68
S2M2-XL [70]	1.03	0.08	0.10	**2.94**	1.63	1.00	–	–	–	–	–	–	406	160	3.56
WAFT-Stereo-L† [103]	0.97	0.07	0.10	4.39	2.15	1.14	1.47	0.95	1.28	1.05	–	–	380	150	2.83
DispViT+ [31]	2.56	0.29	0.12	–	–	–	1.59	1.02	1.51	1.12	–	–	1282	134	6.25
NBS [93]	0.71	**0.04**	0.09	4.22	1.73	1.05	–	–	–	–	–	–	352	323	2.52
StereoViT-M†	0.68	0.10	0.09	–	–	–	1.30	0.84	1.33	1.00	6.07	2.41	176	**41**	**0.96**
StereoViT-L†	**0.64**	**0.04**	**0.08**	3.12	**1.52**	**0.98**	**1.19**	**0.78**	**1.20**	**0.95**	**5.81**	**2.21**	474	97	2.16

sentation for policy learning in Sec. 4.3, and also ablate its design in Sec. 4.4.

Implementation details. Both variants are trained for 200k steps with batch size 128 on 952×1680 crops. We use AdamW [65] with a base learning rate of 10^{-4}, cosine decay, and 5% linear warmup; the pretrained encoder uses $0.2\times$ the base learning rate. Training is performed on 32 NVIDIA GH200 GPUs. Benchmark fine-tuning uses 1456×2520 resolution, 25k steps, batch size 32, and a learning rate of 2×10^{-5}.

4.1. Stereo matching

Setup. For zero-shot generalization, we evaluate the synthetic-only SynLarge checkpoints on the training splits of KITTI [27, 69], ETH3D [84], and Middlebury [83]. For official test-set evaluation, we use the single fine-tuned StereoViT-L checkpoint described in Sec. 3.2. We report BP-x, the percentage of pixels with disparity error exceeding x pixels; D1, the percentage with error exceeding both 3 pixels and 5% of the ground-truth disparity; and EPE, the mean absolute disparity error in pixels. We also report parameter counts and measure latency and peak memory on 544×960 inputs using a single GH200 GPU in FP32. Full benchmark results and qualitative comparisons are provided in the supplementary.

Official benchmarks. StereoViT-L achieves the best or second-best result on every metric reported in Tab. 1. It improves on WAFT-Stereo-L across all reported metrics on KITTI 2012 and KITTI 2015. On ETH3D, it achieves the lowest BP-0.5 without fine-tuning on ETH3D. On Middlebury, it leads on BP-4 and EPE, while S2M2-XL retains an advantage at the stricter BP-2 threshold. On Spring, StereoViT-L achieves the lowest errors on reported aggregate metrics, with StereoViT-M ranking second.

Efficiency. Crucially, these accuracy gains do not compromise efficiency and parameter count does not predict latency. StereoViT-L is the second largest model in Tab. 1 yet runs in 97 ms, accelerating inference by 1.5–4.8× compared to state-of-the-art baseline models. Meanwhile, StereoViT-M emerges as the fastest and most memory-efficient model in our comparison: at just 41 ms and 0.96 GB, it runs $11\times$ faster than FoundationStereo while still securing second place on ETH3D. By discarding cost volumes and iterative updates, StereoViT breaks the traditional trade-off between speed and accuracy, improving both simultaneously.

Table 2. **Zero-shot generalization.** All methods are evaluated on the training splits on all pixels; Middlebury at quarter resolution. Lower is better.

Method	ETH3D		Middlebury		KITTI 2012		KITTI 2015	
	BP-1	EPE	BP-2	EPE	D1	EPE	D1	EPE
RAFT-Stereo [60]	2.60	0.27	10.25	1.21	4.06	0.90	5.43	1.13
UniMatch [116]	6.24	0.42	9.71	1.13	4.91	1.10	5.77	1.21
Selective-IGEV [101]	6.07	1.41	9.29	1.14	5.18	1.09	5.79	1.25
BridgeDepth [30]	1.25	0.23	4.30	0.62	3.33	0.83	4.52	1.08
FoundationStereo [108]	0.65	0.17	2.58	0.49	**2.52**	**0.67**	**2.83**	**0.86**
S2M2-XL [70]	0.58	0.19	2.74	0.43	4.06	0.85	3.09	0.89
WAFT-Stereo-L [103]	0.53	0.15	2.22	0.42	2.83	0.73	3.08	0.90
StereoViT-M	0.54	0.14	2.05	0.36	2.57	0.71	2.97	0.89
StereoViT-L	**0.48**	**0.13**	**1.82**	**0.33**	2.61	0.71	3.23	0.94

Zero-shot generalization. Tab. 2 shows that StereoViT-L achieves the lowest errors on ETH3D and Middlebury among the evaluated methods, reducing Middlebury BP-2 by 18% relative to WAFT-Stereo-L. On KITTI, StereoViT-M ranks second in D1, trailing FoundationStereo by 0.05 and 0.14 percentage points on KITTI 2012 and KITTI 2015, respectively. Scaling from M to L improves performance on ETH3D and Middlebury but not KITTI, indicating that the benefits of increased model capacity are dataset-dependent.

4.2. Robot grasping

Setup. We evaluate on DREDS [20] following D3RoMa [106], using material-randomized scenes contain-

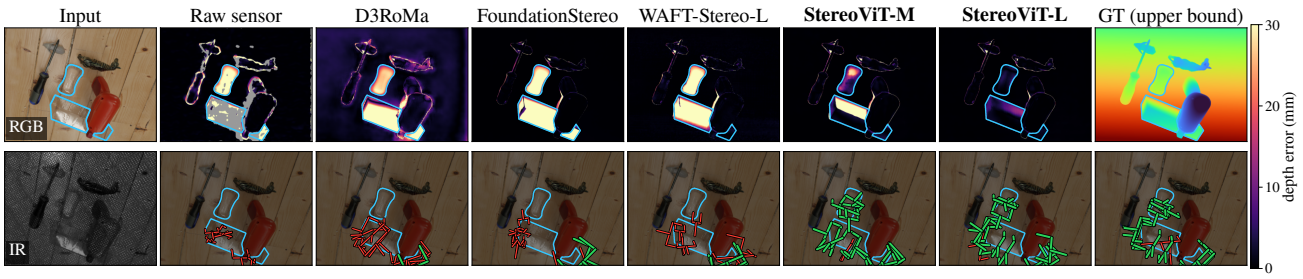


Figure 3. **Qualitative depth and grasping results on DREDS.** Each scene shows depth errors (top; 0–30 mm; gray: missing depth; blue outlines: transparent objects) and EconomicGrasp [111] predictions on transparent objects (bottom; green/red: success/failure under GraspNet evaluation [24]). Inputs show reference RGB and left IR images. Raw sensor denotes simulated D415 depth; GT shows ground-truth depth and grasps from its point cloud.

Table 3. **Depth and grasping on DREDS.** Grasps are predicted by EconomicGrasp [111]. AP_T evaluates transparent objects; SR measures simulated execution success, with bootstrap standard deviations shown as subscripts. †: trained on DREDS; all other stereo models are evaluated without DREDS-specific training.

	Depth		Grasping		
Method	RMSE (m)↓	$\delta_{1.05}$ (%)↑	AP↑	AP_T↑	SR (%)↑
Raw sensor	0.0327	94.98	33.6	3.1	$60.8_{\pm 2.9}$
D3RoMa† [106]	0.0155	95.17	35.9	10.3	$63.1_{\pm 3.0}$
FoundationStereo [108]	0.0160	96.06	42.3	10.0	$69.2_{\pm 3.3}$
WAFT-Stereo-L [103]	0.0125	96.67	43.4	14.7	$68.9_{\pm 2.6}$
StereoViT-M	0.0132	96.30	43.2	13.8	$66.9_{\pm 3.1}$
StereoViT-L	**0.0101**	**97.29**	**43.9**	**16.0**	$\mathbf{72.2}_{\pm 3.2}$
GT depth (upper bound)	–	–	*46.3*	*27.9*	$75.8_{\pm 2.9}$

ing diffuse, specular, and transparent objects. All stereo methods receive the same simulated RealSense D415 infrared pair. Their predicted depth is back-projected into point clouds and passed to a fixed EconomicGrasp detector [111]. All stereo models are evaluated without DREDS-specific training except D3RoMa. We measure depth RMSE and $\delta_{1.05}$, the fraction of pixels within 5% of ground truth, at 640×360 over ground-truth depths between 0.2 and 2 m. For grasping, we report GraspNet-1Billion AP [24] over all objects and transparent objects only (AP_T), assessing collision and force closure against ground-truth meshes. We also execute the top-ranked grasp with a simulated Franka hand and report success rate (SR), with uncertainty estimated by bootstrap resampling. Ground-truth depth provides a reference for the same grasp detector. Further evaluation details are provided in the supplementary.

Results. StereoViT-L achieves the best depth accuracy and the highest grasp AP and mean SR among the evaluated stereo methods (Tab. 3). Overall grasp AP differs only modestly among the strongest baselines, whereas the improvement is more pronounced on transparent objects. Thus, aggregate grasp precision can understate the benefit of improved depth on challenging materials. The qualitative examples in Fig. 3 illustrate this distinction: missing or background-depth estimates misplace grasp contacts, while improved surface recovery supports successful grasps. StereoViT-L also outperforms the DREDS-trained D3RoMa without target-dataset training or changes to the grasp detector, demonstrating transfer to material-diverse grasping. However, transparent objects remain challenging.



Figure 4. **RoboTwin 2.0 tasks.** The six bimanual manipulation tasks used in our simulation experiments, seen from the head camera that provides the policies' visual input.

4.3. Policy learning

4.3.1. Simulated experiments

Setup. We evaluate on six bimanual tasks from RoboTwin 2.0 [16] (Fig. 4), observed by the head camera and a second rendered view with a 6 cm baseline that forms a rectified stereo pair. Each policy is trained on 50 demonstrations per task in clean scenes and evaluated over 100 episodes in clean scenes (*Easy*) and domain-randomized (*Hard*), with distractor clutter and unseen textures and lighting. All methods share the diffusion action model, training recipe, and demonstrations, but differ in their visual inputs and front ends. We compare three groups: (1) *encoders learned from demonstrations*: a ResNet on monocular RGB (DP [19]) or stacked stereo RGB (Stereo-DP), and StereoPolicy [36], which learns stereo fusion on frozen DINOv2 [75] features for each task; (2) *explicit 3D input*: monocular RGB with ground-truth depth (RGBD-DP) and ground-truth point clouds (DP3 [125]); (3) *frozen pretrained features* processed through our adapter: DAv2-B [119] on the left view, and VGGT [99], FoundationStereo [108], and StereoViT on the stereo pair. Feature extraction follows prior work [21, 36, 74]. Full configurations are provided in the supplementary.

Table 4. **Policy learning on RoboTwin 2.0.** All methods follow a shared policy-training recipe and differ in their visual inputs and representations. We report per-task success rates (%) in the Hard and Easy settings. The Mean columns report the average across the six tasks ± standard error, using 600 evaluation episodes per method per setting. (†): Ground-truth depth or point clouds. (*): Our re-implementation of StereoPolicy [36], using a frozen DINOv2 encoder and stereo fusion learned from task demonstrations.

Method	Input	Cans Plasticbox		Burger Fries		Stack Bowls Three		Click Alarmclock		Can Basket		Move Can Pot		Mean	
		Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy	Hard	Easy
Encoders learned from demonstrations															
DP [19]	mono RGB	0	40	0	72	0	63	5	61	0	18	0	39	0.8±0.4	48.8±1.9
Stereo-DP	stereo RGB	0	57	0	80	0	66	7	66	0	25	0	78	1.2±0.4	62.0±1.8
StereoPolicy* [36]	stereo RGB	29	**92**	53	98	0	62	30	**87**	1	42	12	86	20.8±1.5	77.8±1.5
Explicit 3D input															
RGBD-DP	mono RGBD†	0	57	2	83	0	42	4	49	0	25	0	52	1.0±0.4	51.3±1.7
DP3 [125]	point cloud†	3	48	18	72	5	57	14	77	2	67	6	70	8.0±1.1	65.2±1.9
Frozen pretrained representations															
DAv2-B [119]	mono RGB	35	89	49	**99**	7	65	18	76	1	66	7	73	19.5±1.6	78.0±1.5
VGGT [99]	stereo RGB	1	27	25	73	0	38	9	25	0	31	6	77	6.8±1.0	45.2±1.8
FoundationStereo [108]	stereo RGB	13	78	18	83	1	49	12	69	0	62	1	78	7.5±1.0	69.8±1.8
StereoViT-M	stereo RGB	**52**	79	**57**	87	7	66	**48**	79	**15**	**75**	35	**92**	**35.7**±1.8	79.7±1.6
StereoViT-L	stereo RGB	**52**	80	45	84	**10**	**76**	42	86	13	71	**36**	87	33.0±1.8	**80.7**±1.6

Results. StereoViT achieves competitive clean-scene performance while improving robustness under domain randomization, where StereoViT-M reaches 1.7× the success rate of the strongest baseline (Tab. 4).

Additional views and accurate geometry alone do not ensure robustness. Stacking stereo images improves DP's Easy performance but offers little benefit on Hard. Policies supplied with ground-truth depth or point clouds are also less robust than those using StereoViT's frozen features. Thus, the robustness advantage cannot be explained by additional views or depth accuracy alone.

Geometric prediction and feature transfer are distinct capabilities. DAv2-B achieves comparable Easy performance but falls behind on Hard, supporting the value of the full stereo backbone beyond its monocular initialization. The weaker transfer of VGGT and FoundationStereo further shows that geometric pretraining does not guarantee effective frozen policy features. In particular, FoundationStereo's strong disparity accuracy does not translate into comparable downstream robustness.

Pretrained stereo features improve robustness beyond task-specific fusion. StereoPolicy already uses frozen pretrained image features and performs comparably in clean scenes, yet StereoViT-M outperforms it on every Hard task. This supports transferring the pretrained stereo backbone rather than learning fusion from limited task demonstrations in the evaluated configurations. Scaling from M to L brings no consistent policy gain, indicating that greater capacity alone does not ensure better transfer.

4.3.2. Real-world experiments

Setup. We evaluate robustness under unseen backgrounds and object instances on a 7-DoF Franka arm observed by an external ZED Mini stereo camera. The six tasks involve transparent, reflective, and small objects and require precise placement: inserting a straw into a bottle, placing a lid on a pot, pouring particles from a glass into a bowl, placing a cube on a riser, hanging a mug on a stand, and stacking a cube on a block (Fig. 5). For each task, all methods use the same 150 teleoperated demonstrations collected in clean scenes. Evaluation introduces unseen tablecloths and object instances, with 20 rollouts per method and task, matched initial placements across methods, and a 40 s limit per rollout. We report successful object lifting (*Pick*) and full task completion (*Task*) separately, and compare the image-based and feature-based baselines from Tab. 4 with the same augmentation and training recipe. More experiments and de-

Table 5. **Real-robot rollouts in out-of-distribution scenes.** Task success rate (%) over 20 rollouts per method and task, the same 20 initial placements for every method; in parentheses the pick rate (grasping and lifting the target object). SP: StereoPolicy [36]; FS: FoundationStereo [108].

Task	DP [19]	Stereo-DP	SP	FS	StereoViT-M
Insert straw	0 (0)	0 (0)	0 (10)	5 (15)	**30** (90)
Place lid	0 (0)	0 (0)	5 (30)	20 (35)	**65** (80)
Pour glass	0 (0)	0 (0)	35 (45)	90 (90)	**95** (95)
Place riser	0 (0)	0 (0)	0 (10)	15 (30)	**40** (75)
Hang cup	0 (70)	0 (60)	5 (40)	50 (95)	**75** (100)
Stack cube	0 (0)	0 (5)	5 (35)	25 (45)	**40** (65)
Mean	0.0 (11.7)	0.0 (10.8)	8.3 (28.3)	34.2 (51.7)	**57.5** (84.2)

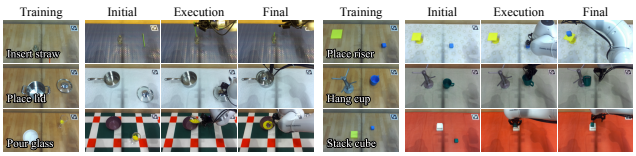


Figure 5. **Real-robot tasks.** For each of the six tasks: a training scene from the demonstrations, and a rollout of StereoViT in an unseen scene (new tablecloth and object instances) at the initial state, during execution, and at the final state.

Table 6. **Ablation of stereo model design.** Each row changes one component of the full model (top). *Attn.*: row (1D) or global (2D) cross-attention in the global pass. *D/U*: global pass on the 2× pooled grid. *Head*: expectation over disparity anchors (Bins) or direct regression. *Data*: our mixture (SynLarge), SceneFlow only, or the zero-shot mixture of FoundationStereo [108] (FSD-ZS). Bad-pixel rates (%, ↓) on all pixels, as in Tab. 2: KITTI D1, ETH3D BP-1, Middlebury-Q BP-2.

	Enc.	Attn.	D/U	Input	Data	Head	Params (M)	KITTI-12	KITTI-15	ETH3D	Midd-Q
Full model	DAv2-B	Row	✓	952×1680	SynLarge	Bins	175.6	**2.58**	**2.99**	1.06	**2.69**
Attention	DAv2-B	Global	✓	952×1680	SynLarge	Bins	175.6	2.93	3.22	1.12	3.32
	DAv2-B	Row	–	952×1680	SynLarge	Bins	167.9	2.98	3.30	1.16	3.46
Encoder	CroCo	Global	✓	952×1680	SynLarge	Bins	102.2	4.06	4.13	1.51	4.68
	DINOv3-B	Row	✓	952×1680	SynLarge	Bins	158.9	2.95	3.37	0.99	3.74
	DAv3-B	Row	✓	952×1680	SynLarge	Bins	176.8	2.96	3.26	**0.97**	3.57
Data / Input	DAv2-B	Row	✓	952×1680	SceneFlow	Bins	175.6	59.0	80.8	99.4	64.3
	DAv2-B	Row	✓	952×1680	FSD-ZS	Bins	175.6	3.65	3.69	1.72	3.77
	DAv2-B	Row	✓	544×960	SynLarge	Bins	175.6	2.81	3.32	1.24	3.29
Head	DAv2-B	Row	✓	952×1680	SynLarge	Regr.	175.1	2.91	3.46	1.42	3.77

Table 7. **Ablation of the policy input.** Mean success (%) ± standard error over six RoboTwin 2.0 tasks using StereoViT-M. Alternative inputs replace the final decoder features; additional inputs concatenate left-view RGB (via ResNet-18), frozen DINOv2 tokens, or stereo features from wrist pairs.

		Alternative input			Additional input		
Setting	Base	Block 2	Block 4	Depth	+ RGB	+ DINOv2	+ Wrist
Easy	79.7±1.6	78.7±1.6	80.2±1.6	55.8±1.7	65.3±1.9	**81.2**±1.5	39.0±1.5
Hard	**35.7**±1.8	26.2±1.6	**35.7**±1.7	18.5±1.4	12.5±1.3	11.2±1.2	8.5±0.9

tails are provided in the supplementary.

Results. Tab. 5 shows that additional views alone do not ensure robust real-world manipulation. Neither the monocular RGB policy nor Stereo-DP completes a single rollout under the scene and object changes; at best they still grasp the mug in hang cup but never hang it. In contrast, policies using pretrained visual features complete tasks, supporting the value of visual priors beyond the task demonstrations. FoundationStereo and StereoViT outperform StereoPolicy on every task, suggesting that stereo-pretrained features provide a more effective starting point than learning stereo fusion from limited demonstrations. StereoViT achieves the highest pick and task success on all six tasks. These results support the transferability of StereoViT's frozen stereo features to robust real-world control, although precise straw insertion remains challenging.

4.4. Ablations

Setup. Tab. 6 evaluates architectural and training variants of StereoViT-M in the zero-shot setting after 50k steps, with batch size 128 and learning rate 10^{-4}. Tab. 7 varies the policy input on the six RoboTwin 2.0 tasks, keeping the action model and training recipe fixed.

Stereo model. Row-restricted cross-attention outperforms 2D attention at matched parameter count, while the anchor-based readout reduces error by 11–29% relative to direct regression. These results support retaining lightweight stereo-specific guidance within the feed-forward model. Pooling the global pass also improves every benchmark, favoring coarse global interaction alongside local attention. Among encoders, DAv2 performs best on KITTI and Middlebury, while DINOv3 [89] and DAv3 [58] achieve slightly lower ETH3D error. DAv2 also outperforms CroCo when both use global cross-attention, although their different capacities prevent isolating the effect of pretraining. Higher input resolution and replacing FSD-ZS with SynLarge each improve all four benchmarks. SceneFlow alone does not transfer to real images, similar to WAFT-Stereo [103] results. Without matching-specific structure, correspondence has to be learned from data, so the model depends on large and diverse training sets.

Policy input. Decoder depth affects robustness more than clean-scene performance: block 2 performs comparably on Easy but is weaker on Hard, whereas block 4 already matches the final output. Thus, the policy benefits from later decoder features, but using the final block is not essential. Replacing features with the same model's predicted depth lowers success in both settings, favoring feature-based conditioning for this policy.

Additional inputs do not improve robustness under the tested fusion scheme. DINOv2 tokens slightly improve Easy success but substantially reduce Hard success, showing that clean-scene gains can conceal weaker robustness. RGB and wrist stereo inputs reduce success even on Easy, so their degradation is not specific to domain randomization. More visual information therefore does not automatically yield more effective policy inputs.

5. Conclusion

We presented StereoViT, a stereo vision transformer that produces accurate disparity and transferable stereo features in a single forward pass. With one checkpoint, it achieves

leading accuracy across five stereo benchmarks while maintaining efficient inference. Its predicted depth improves grasping on transparent objects, while its frozen features improve diffusion-policy robustness under unseen visual variation in both simulator and real scenes. Our policy study focuses on task-specific diffusion policies. Extending StereoViT to vision-language-action models [10, 12] and world-action models [2] is a natural next step. Currently, StereoViT processes stereo pairs independently per frame, leaving temporal consistency across video sequences as an avenue for future work.

Acknowledgement

This work was supported by the Swiss AI Initiative through a grant from the Swiss National Supercomputing Centre (CSCS) under project ID ab036 on the Alps infrastructure; the Swiss National Science Foundation (Advanced Grant 216260: "Beyond Frozen Worlds: Capturing Functional 3D Digital Twins from the Real World"); the European Union's Horizon Europe research and innovation programme under grant agreement No. 101214398 (ELLIOT); and the Horizon Europe programme under grant agreement No. 101298421 (GRAIL). Dingxi Zhang is also supported by the Max Planck ETH Center for Learning Systems (CLS).

References

[1] Ali Abouzeid, Malak Mansour, Qinbo Sun, Zezhou Sun, and Dezhen Song. Geoaware-vla: Implicit geometry aware vision-language-action model. *arXiv preprint arXiv:2509.14117*, 2025. 3

[2] Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. 1, 3, 9

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 3

[4] Kaixin Bai, Huajian Zeng, Lei Zhang, Yiwen Liu, Hongli Xu, Zhaopeng Chen, and Jianwei Zhang. Cleardepth: Efficient stereo perception of transparent objects for robotic manipulation. *arXiv preprint arXiv:2409.08926*, 2024. 3

[5] Kaixin Bai, Lei Zhang, Yiwen Liu, Zhaopeng Chen, and Jianwei Zhang. Stereoanything: Advanced zero-shot stereo imaging for robotic grasp detection with transparent objects. *IEEE Transactions on Cybernetics*, 56(5), 2026. 3

[6] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: A large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63:1–11, 2020. 4

[7] Vladislav Bargatin, Alexander Yakovenko, Khaled Abud, and Dmitriy Vatolin. Freeflow: A bias-free hierarchical transformer for optical flow estimation. In *European Conference on Computer Vision*, pages 188–206. Springer, 2026. 2, 3, 15

[8] Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In *CVPR*, pages 1013–1027, 2025. 2, 16

[9] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018, 2021. 3

[10] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 1, 3, 9

[11] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1

[12] Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 3, 9

[13] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, pages 611–625, 2012. 4

[14] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 4

[15] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 2

[16] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, 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. 2, 6

[17] Ziyang Chen, Wei Long, He Yao, Yongjun Zhang, Bingshu Wang, Yongbin Qin, and Jia Wu. Mocha-stereo: Motif channel attention network for stereo matching. In *CVPR*, pages 27768–27777, 2024. 2

[18] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. In *CVPR*, pages 6273–6282, 2025. 2, 16

[19] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 1, 4, 6, 7

[20] Qiyu Dai, Jiyao Zhang, Qiwei Li, Tianhao Wu, Hao Dong, Ziyuan Liu, Ping Tan, and He Wang. Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects. In *ECCV*, pages 374–391, 2022. 1, 2, 3, 5, 17

[21] Shengliang Deng, Mi Yan, Yixin Zheng, Jiayi Su, Wenhao Zhang, Yitao Zeng, Xiaoguang Zhao, Heming Cui, Zhizheng Zhang, and He Wang. Stereovla: Enhancing vision-language-action models with stereo vision. In *Robotics: Science and Systems (RSS)*, 2026. arXiv:2512.21970. 2, 3, 6, 19

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[23] Hongjie Fang, Hao-Shu Fang, Sheng Xu, and Cewu Lu. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *IEEE Robotics and Automation Letters*, 7(3):7383–7390, 2022. 1, 3

[24] Hao-Shu Fang, Minghao Gou, Chenxi Wang, and Cewu Lu. Robust grasping across diverse sensor qualities: The graspnet-1billion dataset. *The International Journal of Robotics Research*, 2023. 6, 17

[25] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023. 3

[26] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17461–17474. IEEE, 2025. 3

[27] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 4, 5

[28] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning (CoRL)*, pages 694–710, 2023. 3

[29] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching. In *CVPR*, pages 5459–5469, 2024. 2

[30] Tongfan Guan, Jiaxin Guo, Chen Wang, and Yun-Hui Liu. Bridgedepth: Bridging monocular and stereo reasoning with latent alignment. In *ICCV*, pages 27681–27691, 2025. 2, 5, 16

[31] Tongfan Guan, Jiaxin Guo, Tianyu Huang, Jinhu Dong, Chen Wang, and Yun-Hui Liu. Dispvit: Direct stereo disparity regression with a single-stream vision transformer. In *The Fourteenth International Conference on Learning Representations*, 2026. 2, 5, 16

[32] Weiyu Guo, Zhaoshuo Li, Yongkui Yang, Zheng Wang, Russell H. Taylor, Mathias Unberath, Alan Yuille, and Yingwei Li. Context-enhanced stereo transformer. In *ECCV*, pages 263–279, 2022. 2

[33] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, pages 3273–3282, 2019. 2

[34] Xianda Guo, Chenming Zhang, Youmin Zhang, Ruilin Wang, Dujun Nie, Wenzhao Zheng, Matteo Poggi, Hao Zhao, Mang Ye, Qin Zou, et al. Stereo anything: Unifying zero-shot stereo matching with large-scale mixed data. *arXiv preprint arXiv:2411.14053*, 2024. 2

[35] Beining Han, Yu-Wei Chao, Erwin Coumans, Clemens Eppner, Jia Deng, Stan Birchfield, and Adithyavairavan Murali. Graspgen-x: Cross-embodiment 6-dof diffusion-based grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20878–20889, 2026. 3

[36] Evans Han, Yunfan Jiang, Yingke Wang, Haoyue Xiao, Huang Huang, Jianwen Xie, Jiajun Wu, Li Fei-Fei, and Ruohan Zhang. Stereopolicy: Improving robotic manipulation policies via stereo perception. *arXiv preprint arXiv:2605.09989*, 2026. 2, 3, 6, 7, 19, 20

[37] Jisang Han, Seonghu Jeon, Jaewoo Jung, René Zurbrügg, Honggyu An, Tifanny Portela, Marco Hutter, Marc Pollefeys, Seungryong Kim, and Sunghwan Hong. Geometric action model for robot policy learning. *arXiv preprint arXiv:2606.17046*, 2026. 3

[38] Kush Hari, Justin Kerr, Nidhya Shivakumar, Samarth Mahapatra, Carmelo Sferrazza, Jiahui Lei, Jitendra Malik, C. Karen Liu, Ken Goldberg, and Angjoo Kanazawa. EyeRobot 2.0: Active gaze for precise manipulation without wrist cameras. *arXiv preprint arXiv:2610.03710*, 2026. 2, 3

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[40] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. 2

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[42] Junwen Huang, Jizhong Liang, Jiaqi Hu, Martin Sundermeyer, Peter K. T. Yu, Nassir Navab, and Benjamin Busam. Xyz-ibd: High-precision bin-picking dataset for object 6d pose estimation capturing real-world industrial complexity. *arXiv preprint arXiv:2506.00599*, 2025. 16

[43] Byungwoo Jeon, Sunghwan Hong, Dohyeon Kim, Jonghoon Lee, Jungwoo Park, Inhee Lee, Doosan Baek, Jinil Kim, Taehyun Lim, Hanbyul Joo, et al. Learning the 4d world through point trajectory forecasting. *arXiv preprint arXiv:2610.28594*, 2026. 3

[44] Yueru Jia, Jiaming Liu, Sixiang Chen, Chenyang Gu, Zhilue Wang, Longzan Luo, Lily Lee, Pengwei Wang, Zhongyuan Wang, Renrui Zhang, and Shanghang Zhang. Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation. In *CVPR*, 2025. 3

[45] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *CVPR*, pages 21857–21867, 2025. 2, 16

[46] Junpeng Jing, Ye Mao, Anlan Qiu, and Krystian Mikolajczyk. Match stereo videos via bidirectional alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2026. 16, 17

[47] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, pages 1949–1974, 2024. 3

[48] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 3

[49] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, pages 66–75, 2017. 2, 3

[50] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024. 16, 17

[51] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024. 1, 3

[52] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026. 3

[53] Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *IEEE Robotics and Automation Letters*, 2025. 3

[54] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025. 3

[55] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *CVPR*, pages 16263–16272, 2022. 2, 4

[56] Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models. In *NeurIPS*, 2025. 3

[57] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *ICCV*, pages 6197–6206, 2021. 2

[58] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 3, 8

[59] Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, and Bo Zhao. Evo-0: Vision-language-action model with implicit spatial understanding. *arXiv preprint arXiv:2507.00416*, 2025. 3

[60] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, pages 218–227, 2021. 2, 5, 16

[61] I-Chun Arthur Liu, Jason Chen, Gaurav S. Sukhatme, and Daniel Seita. Exstereo: Lifting 2d vision-language-action models to 3d with explicit stereo representations. *arXiv preprint arXiv:2610.04805*, 2026. 2, 3

[62] Jiangyuan Liu, Hongxuan Ma, Yuxin Guo, Yuhao Zhao, Chi Zhang, Wei Sui, and Wei Zou. Monocular depth estimation and segmentation for transparent object with iterative semantic and geometric fusion. In *ICRA*, 2025. arXiv:2502.14616. 3

[63] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002, 2021. 2, 3

[64] Zhidan Liu, Chengtang Yao, Jiaxi Zeng, Yuwei Wu, and Yunde Jia. Multi-label stereo matching for transparent scene depth estimation. *arXiv preprint arXiv:2505.14008*, 2025. 3

[65] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[66] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? In *NeurIPS*, 2023. 1, 3

[67] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 2, 4

[68] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *CVPR*, pages 4981–4991, 2023. 2, 4

[69] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015. 2, 4, 5, 15

[70] Junhong Min, Youngpil Jeon, Jimin Kim, and Minyong Choi. S2m2: Scalable stereo matching model for reliable depth estimation. In *ICCV*, pages 26729–26739, 2025. 2, 5, 16

[71] Junhong Min, Jimin Kim, Cheol-Hui Min, Minwook Kim, Youngpil Jeon, and Minyong Choi. Depthfocus: Con-

trollable depth estimation for see-through scenes. *arXiv preprint arXiv:2511.16993*, 2025. 3

[72] Adithyavairavan Murali, Balakumar Sundaralingam, Yu-Wei Chao, Wentao Yuan, Jun Yamada, Mark Carlson, Fabio Ramos, Stan Birchfield, Dieter Fox, and Clemens Eppner. Graspgen: A diffusion-based framework for 6-dof grasping with on-generator training. In *ICRA*, 2026. arXiv:2507.13097. 3

[73] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Proceedings of The 6th Conference on Robot Learning (CoRL)*, pages 892–909, 2022. 1, 3

[74] Zehao Ni, Yonghao He, Lingfeng Qian, Jilei Mao, Fa Fu, Wei Sui, Hu Su, Junran Peng, Zhipeng Wang, and Bin He. Vo-dp: Semantic-geometric adaptive diffusion policy for vision-only robotic manipulation. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9975–9982. IEEE, 2026. 1, 3, 6, 19

[75] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6, 19

[76] Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025. 3

[77] Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. Tartanground: A large-scale dataset for ground robot perception and navigation. In *IROS*, 2025. 4

[78] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4

[79] Quanhao Qian, Guoyang Zhao, Gongjie Zhang, Jiuniu Wang, Ran Xu, Junlong Gao, and Deli Zhao. Gp3: A 3d geometry-aware policy with multi-view images for robotic manipulation. *arXiv preprint arXiv:2509.15733*, 2025. 3

[80] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025. 1, 3

[81] Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Booster: A benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4, 15, 16

[82] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *ICRA*, pages 3634–3642, 2020. 1, 3

[83] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR)*, pages 31–42, 2014. 2, 4, 5

[84] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 2, 4, 5, 16

[85] Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, pages 724–748, 2024. 1, 3

[86] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *CVPR*, pages 13906–13915, 2021. 2

[87] Jun Shi, A Yong, Yixiang Jin, Dingzhe Li, Haoyu Niu, Zhezhu Jin, and He Wang. Asgrasp: Generalizable transparent object reconstruction and 6-dof grasp detection from rgb-d active stereo camera. In *2024 IEEE international conference on robotics and automation (ICRA)*, pages 5441–5447. IEEE, 2024. 3

[88] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025. 1, 3

[89] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 8

[90] Zijian Song, Qichang Li, Jiawei Zhou, Zhenlong Yuan, Tianshui Chen, Liang Lin, and Guangrun Wang. Robotic manipulation is vision-to-geometry mapping ($f(v) \rightarrow g$): Vision-geometry backbones over language and video models. *arXiv preprint arXiv:2604.12908*, 2026. 3

[91] Guoheng Sun, Tingting Du, Kaixi Feng, Chenxiang Luo, Xingguo Ding, Zheyu Shen, Ziyao Wang, Yexiao He, and Ang Li. Rocket: Residual-oriented multi-layer alignment for spatially-aware vision-language-action models. *arXiv preprint arXiv:2602.17951*, 2026. 3

[92] Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*, 2025. 1, 3

[93] Vage Taamazyan, Zhuowen Shen, Stefan Hinterstoisser, Alberto Dall'Olio, Agastya Kalra, Aarrushi Shandilya, Xin Li, Wenping Wang, and Kartik Venkataraman. Nbs: No bias stereo. *arXiv preprint arXiv:2608.28933*, 2026. 2, 3, 5, 16

[94] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field

transforms for optical flow. In *ECCV*, pages 402–419, 2020. 3, 15

[95] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. 17

[96] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *CVPR*, pages 8942–8952, 2021. 4

[97] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *CVPR Workshops*, pages 2038–2041, 2018. 4

[98] Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *ICCV*, pages 15964–15973, 2021. 3, 4, 17, 18

[99] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 3, 6, 7

[100] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IROS*, pages 4909–4916, 2020. 4

[101] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *CVPR*, pages 19701–19710, 2024. 2, 5, 16

[102] Yihan Wang and Jia Deng. Waft: Warping-alone field transforms for optical flow. *arXiv preprint arXiv:2506.21526*, 2025. 4

[103] Yihan Wang and Jia Deng. Waft-stereo: Warping-alone field transforms for stereo matching. In *European Conference on Computer Vision*, pages 210–228. Springer, 2026. 2, 4, 5, 6, 8, 16

[104] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *ECCV*, pages 36–54, 2024. 4, 15

[105] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, pages 10481–10497, 2026. 3

[106] Songlin Wei, Haoran Geng, Jiayi Chen, Congyue Deng, Wenbo Cui, Chengyang Zhao, Xiaomeng Fang, Leonidas Guibas, and He Wang. D^3roma: Disparity diffusion-based depth sensing for material-agnostic robotic manipulation. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, pages 4944–4966, 2024. 3, 5, 6, 17

[107] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pretraining for stereo matching and optical flow. In *ICCV*, pages 17969–17980, 2023. 2, 3, 5

[108] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *CVPR*, pages 5249–5260, 2025. 2, 3, 4, 5, 6, 7, 8, 16

[109] Bowen Wen, Shaurya Dewan, and Stan Birchfield. Fast-foundationstereo: Real-time zero-shot stereo matching. In *CVPR*, 2026. arXiv:2512.11130. 16

[110] Huimin Wu, Kwang-Ting Cheng, Stephen Lin, and Zhirong Wu. A study of finetuning video transformers for multi-view geometry tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10646–10654, 2026. 2

[111] Xiao-Ming Wu, Jia-Feng Cai, Jian-Jian Jiang, Dian Zheng, Yi-Lin Wei, and Wei-Shi Zheng. An economic framework for 6-dof grasp detection. In *European Conference on Computer Vision*, pages 357–375. Springer, 2024. 1, 3, 4, 6, 17, 18

[112] Chongkai Xu, Hongyang Li, Shuo Cheng, Jun Hu, Heng Fan, Zhiwei Feng, and Si Liu. Action-geometry prediction with 3d geometric prior for bimanual manipulation. *arXiv preprint arXiv:2602.23814*, 2026. 3

[113] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, pages 21919–21928, 2023. 2, 16

[114] Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. Igev++: Iterative multi-range geometry encoding volumes for stereo matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[115] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, pages 1959–1968, 2020. 2

[116] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 5, 16

[117] David Yan, Alexander Raistrick, and Jia Deng. What makes good synthetic training data for zero-shot stereo matching? *arXiv preprint arXiv:2504.16930*, 2025. 4

[118] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5510–5519. IEEE, 2019. 4

[119] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, pages 21875–21911, 2024. 2, 3, 4, 6, 7, 15, 20

[120] Rujia Yang, Geng Chen, Chuan Wen, and Yang Gao. Fp3: A 3d foundation policy for robotic manipulation. *arXiv preprint arXiv:2503.08950*, 2025. 3

[121] Yixuan Yang, Sha Zhang, Rui Li, Zhenfei Yin, Xinzhu Ma, Yiran Qin, Lei Bai, Xudong Xu, Shilin Shan, Wangmeng Zuo, Yanyong Zhang, Wanli Ouyang, Feng Zheng, Shixiang Tang, and Dongzhan Zhou. Trans2occ: Voxel occupancy estimation and grasp for transparent objects from simulation to reality. *arXiv preprint arXiv:2606.01777*, 2026. 3

[122] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026. 1, 3

[123] Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Zhuoguang Chen, Tao Jiang, and Hang Zhao. Depthvla: Enhancing vision-language-action models with depth-aware spatial reasoning. *arXiv preprint arXiv:2510.13375*, 2025. 1, 3

[124] Yanjie Ze, Zixuan Chen, Wenhao Wang, Tianyi Chen, Xialin He, Ying Yuan, Xue Bin Peng, and Jiajun Wu. Generalizable humanoid manipulation with 3d diffusion policies. *arXiv preprint arXiv:2410.10803*, 2024. 3

[125] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Robotics: Science and Systems (RSS)*, 2024. 3, 6, 7, 19

[126] Dingxi Zhang, Fangjinhua Wang, Marc Pollefeys, and Haofei Xu. Megaflow: Zero-shot large displacement optical flow. In *European Conference on Computer Vision*, pages 343–362. Springer, 2026. 2

[127] Ruowen Zhao, Bangguo Li, Zuyan Liu, Yinan Liang, Junliang Ye, Fangfu Liu, Diankun Wu, Zhengyi Wang, Xumin Yu, Yongming Rao, et al. Gem: Generative supervision helps embodied intelligence. In *European Conference on Computer Vision*, pages 618–636. Springer, 2026. 1, 3

[128] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023. 1

[129] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. In *ICML*, 2024. 3

[130] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025. 3

[131] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. 3

StereoViT: Stereo Vision Transformer for Robust Robot Manipulation

Supplementary Material

In this supplementary material, we provide additional quantitative and qualitative results and additional implementation details that complement the main paper.

6. Model and training details

Architecture. We extract features from encoder layers $\{3, 6, 9, 12\}$ for ViT-B (StereoViT-M) and $\{5, 12, 18, 24\}$ for ViT-L (StereoViT-L), concatenate them, and linearly project them to the decoder width C. Both views are processed in a single batch by the shared encoder. The decoder operates at $1/14$ input resolution. A 952×1680 input produces $68 \times 120 = 8{,}160$ tokens per view. At this resolution, each window pass partitions the tokens into a 4×4 grid of windows, each containing 17×30 tokens. The global pass operates on the $2\times$ downsampled 34×60 grid, with cross-attention restricted to corresponding epipolar rows.

The disparity head comprises a 3×3 convolution, a 1×1 convolution producing 4096 channels, and a 1×1 projection producing logits over $M = 128$ disparity anchors. A transposed convolution with kernel size and stride 14 predicts nine convex-upsampling weights [94] and three mixture-of-Laplace parameters at each output pixel. The encoder is initialized from Depth Anything V2 [119], and the decoder from CroCo-pretrained FreeFlow [7]. The feature projection and disparity head are trained from scratch.

Supervision. Following SEA-RAFT [104], we supervise the full-resolution disparity with a mixture-of-Laplace (MoL) negative log-likelihood. At each pixel, the head predicts the disparity $\hat{d}$, two mixture logits, and a log-scale γ. Softmax normalization converts the logits into mixture weights α_1 and α_2. Both Laplace components are centered at $\hat{d}$: one has a fixed unit scale $\beta_1 = 1$, while the other has scale $\beta_2 = \exp(\gamma)$, with γ clamped to $[0, 10]$. Writing $e(p) = |d_{\text{gt}}(p) - \hat{d}(p)|$, the loss is

$$\mathcal{L}_{\text{MoL}} = -\frac{1}{|\Omega|} \sum_{p \in \Omega} \log \left[\sum_{i=1}^{2} \frac{\alpha_i(p)}{2\beta_i(p)} \exp\left(-\frac{e(p)}{\beta_i(p)} \right) \right], \tag{S1}$$

where Ω contains pixels with valid ground-truth disparity. The learned-scale component accommodates large residuals, while the unit-scale component retains an L_1-like penalty. The mixture parameters are used only for training.

Checkpoints. We will release four checkpoints: a synthetic-only and a benchmark-fine-tuned checkpoint for each of StereoViT-M and StereoViT-L. The synthetic-only checkpoints, trained on SynLarge at 952×1680, are used for zero-shot stereo evaluation (Tab. 2) and all robotics experiments (Tab. 3-5). The benchmark checkpoints are fine-tuned at 1456×2520 following Sec. 3.2 and used for the official results in Tab. 1.

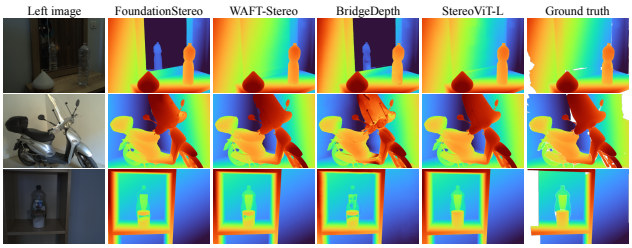


Figure S1. Qualitative zero-shot results on Booster dataset [81]

Table S1. Architectural configurations and parameter counts for StereoViT variants.

Component	StereoViT-M	StereoViT-L
Encoder	DAv2 ViT-B	DAv2 ViT-L
Blocks / Dim	12 / 768	24 / 1024
Projection	$4 \times 768 \to 384$	$4 \times 1024 \to 512$
Decoder	6 blocks, 6 heads	8 blocks, 8 heads
Hidden Dim	384	512
Parameters	175.6M	474M

7. Additional stereo results

Evaluation metrics. For our local evaluations, let Ω_i denote the pixels in image i with valid ground-truth disparity, restricted to non-occluded pixels when stated. Given ground truth $d_{i,p}$ and prediction $\hat{d}_{i,p}$, the absolute disparity error is $e_{i,p} = |\hat{d}_{i,p} - d_{i,p}|$. Over N images, we compute

$$\text{EPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} e_{i,p}, \tag{S2}$$

$$\text{BP-}x = \frac{100}{N} \sum_{i=1}^{N} \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} \mathbf{1}[e_{i,p} > x], \tag{S3}$$

$$\text{D1} = 100 \, \frac{\sum_{i=1}^{N} \sum_{p \in \Omega_i} \mathbf{1}[e_{i,p} > 3 \wedge e_{i,p} > 0.05 d_{i,p}]}{\sum_{i=1}^{N} |\Omega_i|}. \tag{S4}$$

EPE is measured in pixels at the evaluation resolution; BP-x and D1 are percentages. EPE and BP-x are averaged per image, whereas D1 pools valid pixels across images, following the KITTI 2015 protocol [69]. We apply the same D1 criterion to KITTI in the zero-shot evaluation.

Extended zero-shot comparison. Tab. S2 extends Tab.2 with additional stereo models. We use released baseline

Table S2. **Extended zero-shot comparison.** Evaluation on the training splits using all valid ground-truth pixels; Middlebury is evaluated at quarter resolution. BP-x and D1 are reported in %, and EPE in pixels. Latency and peak memory are measured in FP32 on a 544×960 input using one GH200 GPU, following Tab. 1. Lower is better.

Method	ETH3D		Middlebury		KITTI 2012		KITTI 2015		Params	Latency	Memory
	BP-1	EPE	BP-2	EPE	D1	EPE	D1	EPE	(M)	(ms)	(GB)
Cost volume and iterative refinement											
RAFT-Stereo [60]	2.60	0.27	10.25	1.21	4.06	0.90	5.43	1.13	11	139	1.63
IGEV [113]	4.06	0.33	8.88	1.05	4.73	1.03	5.79	1.21	13	141	1.00
Selective-IGEV [101]	6.07	1.41	9.29	1.14	5.18	1.09	5.79	1.25	13	199	1.00
UniMatch [116]	6.24	0.42	9.71	1.13	4.91	1.10	5.77	1.21	7	53	1.56
MonSter [18]	1.33	0.22	5.53	0.62	2.90	0.73	3.23	0.88	389	232	2.75
DEFOM-Stereo [45]	2.37	0.34	5.52	0.62	3.43	0.84	4.77	1.06	383	153	3.23
StereoAnywhere [8]	2.41	0.33	5.62	0.73	3.38	0.87	3.74	0.97	347	340	4.04
BridgeDepth [30]	1.25	0.23	4.30	0.62	3.33	0.83	4.52	1.08	346	97	1.89
FoundationStereo [108]	0.65	0.17	2.58	0.49	**2.52**	**0.67**	**2.83**	**0.86**	375	469	4.68
Fast-FoundationStereo [109]	1.75	0.25	4.42	0.75	2.68	0.70	3.40	0.92	18	85	3.15
S2M2-XL [70]	0.58	0.19	2.74	0.43	4.06	0.85	3.09	0.89	406	160	3.56
Attention or warping without iterative cost filtering											
WAFT-Stereo-L [103]	0.53	0.15	2.22	0.42	2.83	0.73	3.08	0.90	380	150	2.83
DispViT+ [31]	0.57	0.15	3.31	0.52	2.93	0.73	3.09	0.89	1282	134	6.25
StereoViT-M	0.54	0.14	2.05	0.36	2.57	0.71	2.97	0.89	176	**41**	**0.96**
StereoViT-L	**0.48**	**0.13**	**1.82**	**0.33**	2.61	0.71	3.23	0.94	474	97	2.16

checkpoints and their official inference implementations, with a shared evaluation pipeline for ground-truth loading, validity masks, and metric aggregation. Efficiency is measured using the FP32 protocol of Tab. 1. StereoViT-L retains the lowest errors on ETH3D and Middlebury, while StereoViT-M has the lowest latency and peak memory among the compared methods. NBS [93] is excluded because it's not open-sourced.

Additional qualitative results. Fig. S1 compares predictions on Booster scenes with transparent and specular surfaces [81]. We further apply StereoViT-L zero-shot to SouthKensington SV [46] and DROID [50], covering indoor/outdoor scenes and robot manipulation settings, respectively (Fig. S2). Fig. S3 compares back-projected point clouds on ETH3D [84] and XYZ-IBD [42]. On the glass doors, FoundationStereo, WAFT-Stereo-L and S2M2-XL see through the glass and leave large holes, whereas StereoViT-L reconstructs the doors as planar surfaces. In the XYZ-IBD bins of reflective metal parts, WAFT-Stereo-L and S2M2-XL degrade into a smeared surface, while StereoViT-L preserves the individual rods and flanges. The predicted depth is back-projected using the camera calibration, with only points at depth discontinuities removed. The examples show coherent reconstructions of thin and repetitive structures, distant objects, and challenging surfaces such as glass and cloth.

Efficiency in bfloat16. Tab. S3 complements the FP32 comparison with bfloat16 inference, using the same zero-shot checkpoints and re-evaluating accuracy in this precision. StereoViT-M achieves the lowest latency and peak memory, while StereoViT-L obtains the lowest ETH3D BP-0.5 among the evaluated methods. At similar counted FLOPs, StereoViT-M runs 3.3× faster than the real time stereo approach Fast-FoundationStereo, and StereoViT-L runs 3.2× faster than WAFT-Stereo-L. For both StereoViT variants, switching from FP32 to bfloat16 changes the metrics in Tab. 1 by at most 0.01 px for EPE and 0.01 percentage points for error rates.

Table S3. **Runtime and memory efficiency in `bfloat16`.** Latency and peak GPU memory are measured on a 544×960 input using a single GH200 GPU. FLOPs are counted via `torch.utils.flop_counter`. Accuracy is evaluated on the ETH3D training split over all valid pixels.

Method	Params (M)	TFLOPs (↓)	Latency (ms ↓)	Mem. (GB ↓)	BP-0.5 (% ↓)
FoundationStereo [108]	375	27.50	361	3.70	2.96
Fast-FoundationStereo [109]	18	**1.80**	119	8.80	5.78
S2M2-XL [70]	406	12.93	153	2.78	4.05
BridgeDepth [30]	346	2.68	94	1.71	5.31
WAFT-Stereo-L [103]	380	5.83	142	2.84	2.00
StereoViT-M	176	1.95	**36**	**0.53**	1.74
StereoViT-L	474	5.81	45	1.11	**1.48**

Official leaderboards. Fig. S8 shows the public leaderboards at submission time, where StereoViT ranks first on KITTI 2012 and 2015, ETH3D (BP-2), Middlebury (BP-4) and Spring. All metrics remain available on the benchmark websites.

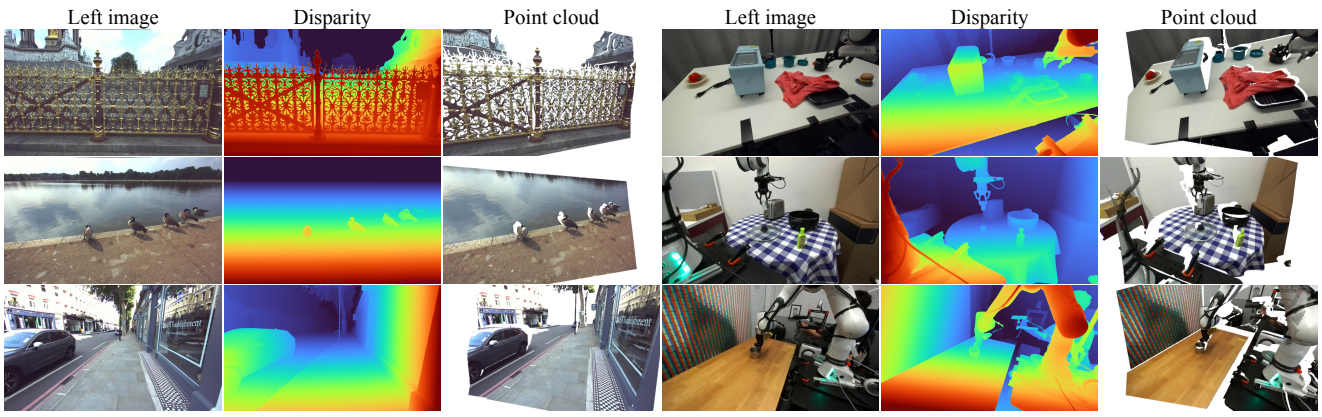


Figure S2. **In the wild results.** SouthKensington SV [46] (left) and DROID [50] (right). Each example shows the left image, predicted disparity, and reconstructed point cloud rendered from a novel viewpoint.

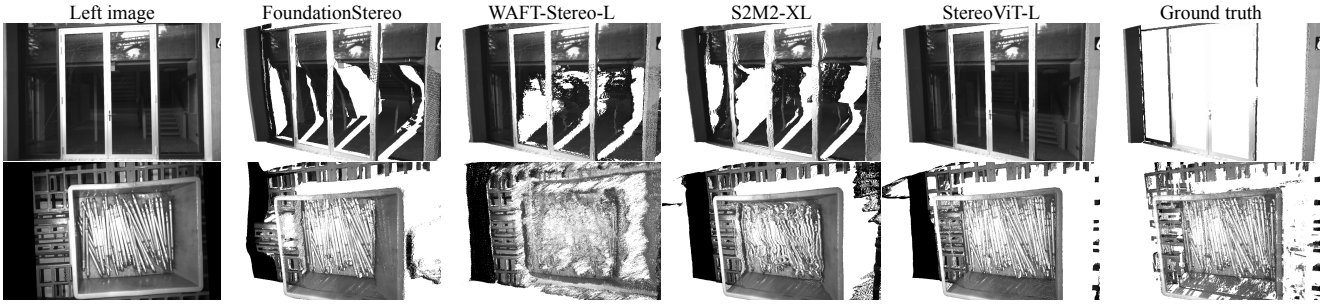


Figure S3. **Zero-shot point clouds on ETH3D and XYZ-IBD**. All methods are back-projected with the dataset calibration and shown from the same novel viewpoint.

8. Grasping on DREDS

Protocol. We evaluate on DREDS-CatNovel [20], using eight views from each of 45 scene layouts derived from GraspNet-1Billion [24], for a total of 360 stereo observations. Object materials are randomized across observations among diffuse, specular, and transparent categories. Each observation includes a simulated RealSense D415 active-IR stereo pair (1280×720, 55 mm baseline), simulated sensor depth, ground-truth depth, instance masks, and object material labels. Object meshes and poses are obtained from the GraspNet-1Billion annotations.

All stereo methods receive the same IR images. Baselines use their released checkpoints and default inference settings, while StereoViT uses the SynLarge-trained checkpoints described in Sec. 3.2. D3RoMa [106] uses a checkpoint trained on DREDS; the remaining stereo methods receive no DREDS-specific training. We compute depth RMSE and $\delta_{1.05}$ at 640×360, restricting evaluation to pixels with ground-truth depth in $[0.2, 2]$ m. Metrics are reported over all valid pixels and separately over object pixels of each material class.

Each depth map is back-projected using the DREDS camera intrinsics and cropped to the workspace with the official GraspNet test-time mask. The resulting point cloud is passed to a fixed pretrained EconomicGrasp [111] or GSNet [98] detector. We evaluate predicted grasps using the GraspNet-1Billion protocol: non-maximum suppression, collision checking against the ground-truth objects and table, and force-closure evaluation. AP averages Precision@k over $k = 1, \ldots, 50$ and friction coefficients $\mu \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$. For each material class, AP is computed using grasps assigned to objects of that class and averaged over views containing at least one such object. A view containing the material but no predicted grasp for that class contributes zero.

For simulated grasp success rate (SR), we retain grasps whose approach direction lies within 30° of vertical, following D3RoMa [106]. We additionally reject grasps that cause the Franka hand to collide with the perceived point cloud of other objects, accounting for differences between the Franka hand and GraspNet's gripper model. The highest-confidence remaining grasp is executed in MuJoCo [95] using a floating Franka hand. After closing, the hand is lifted by 15 cm; a trial succeeds if the object rises by at least 10 cm and remains held. We report the standard deviation of SR over 10,000 bootstrap resamples of the 45 scenes, retaining all eight views of each sampled scene.

Per-material results. Tab. S4 shows that the zero-shot

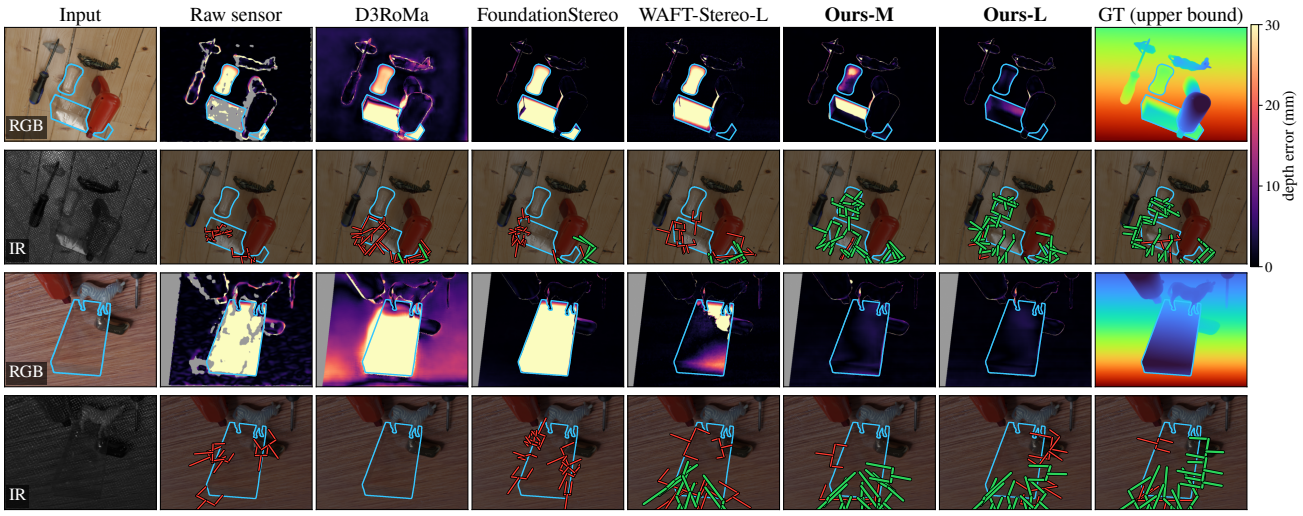


Figure S4. **Additional qualitative depth and grasping results on DREDS-CatNovel.**

stereo methods achieve similar depth accuracy and grasp AP on specular and diffuse objects. Larger differences emerge on transparent objects, where StereoViT-L reduces RMSE by 23% and improves $\delta_{1.05}$ by 9.5 percentage points relative to WAFT-Stereo-L, the strongest baseline on both depth metrics. These gains are accompanied by the highest transparent-object grasp AP. The material breakdown therefore reveals a benefit on challenging surfaces that is less apparent in the aggregate AP reported in Tab. 3. Additional qualitative results are provided in Fig. S4.

Second grasp detector. With GSNet [98], StereoViT-L again achieves the highest AP and AP_T among the evaluated stereo methods (Tab. S5), extending the grasp-precision gains observed with EconomicGrasp [111] to a second detector. StereoViT-M attains the highest observed SR. The different rankings highlight that analytic precision over a ranked grasp set and the success of a single executed grasp measure complementary aspects of downstream performance.

9. Policy Learning Details

Evaluation protocol. The six RoboTwin 2.0 tasks span object transfer, placement, stacking, and button pressing under a common stereo observation protocol, including Place Cans Plasticbox, Place Burger Fries, Stack Bowls Three, Click Alarmclock, Place Can Basket, Move Can Pot. Every method observes the same head-mounted stereo pair and is trained on the same 50 demonstrations per task, collected in clean scenes. We evaluate in clean scenes (*Easy*) and in domain-randomized scenes with cluttered tables, random textures, and varied lighting (*Hard*), which are never seen during training. For each task and setting, every policy is evaluated on the same 100 episodes: the first 100 scene seeds that the scripted expert solves, which fixes the initial object placement, clutter, and appearance across methods. Each episode runs in its own simulator instance with its own policy server.

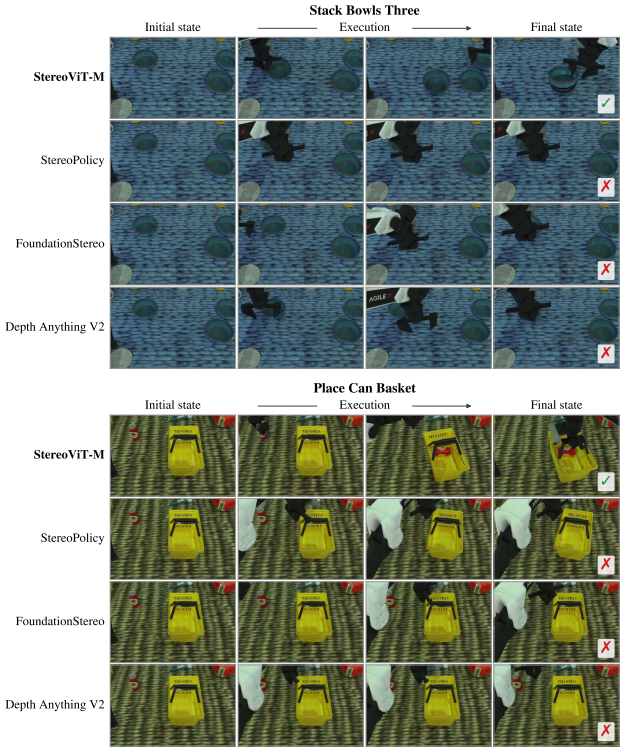


Figure S5. **Qualitative comparison in domain-randomized scenes**. For each task, all policies are evaluated on the same episode (identical scene seed, object placement, clutter, textures and lighting). Columns show the initial state, two intermediate frames taken at the same time steps for every policy, and each policy's final state (task completed or step limit reached); ✓/✗ mark each episode's outcome.

Table S4. **Depth accuracy and grasp AP by material on DREDS-CatNovel.** Depth metrics are evaluated on object pixels of each material class; grasp AP uses EconomicGrasp. RMSE is reported in mm and $\delta_{1.05}$ in %. †: trained on DREDS; other stereo methods are evaluated without DREDS-specific training.

Depth source	Transparent			Specular			Diffuse		
	RMSE↓	$\delta_{1.05}$↑	AP↑	RMSE↓	$\delta_{1.05}$↑	AP↑	RMSE↓	$\delta_{1.05}$↑	AP↑
Raw sensor (sim. D415)	75.9	37.2	3.1	40.5	90.6	23.3	36.6	93.3	26.0
D3RoMa†	45.5	53.2	10.3	14.8	93.3	24.2	14.0	94.7	26.9
FoundationStereo	59.4	46.8	10.0	7.1	98.8	29.7	6.4	99.1	30.7
WAFT-Stereo-L	44.2	56.9	14.7	7.5	98.9	29.8	7.4	98.9	30.3
StereoViT-M	47.0	52.8	13.8	7.5	98.8	30.2	7.1	98.8	30.5
StereoViT-L	34.0	66.4	16.0	7.1	98.9	30.0	6.7	98.9	30.7
GT depth	–	–	*27.9*	–	–	*29.4*	–	–	*29.8*

Table S5. **Grasping with GSNet on DREDS-CatNovel.** We use the same depth predictions and evaluation protocol as in Tab. 3. AP_T evaluates transparent objects; SR subscripts report bootstrap standard deviations. †: trained on DREDS.

Depth source	AP↑	AP_T↑	SR (%)↑
Raw sensor (sim. D415)	19.3	2.0	$44.7_{\pm 3.1}$
D3RoMa†	20.5	5.6	$43.3_{\pm 3.1}$
FoundationStereo	23.7	6.0	$51.7_{\pm 3.3}$
WAFT-Stereo-L	24.6	8.5	$50.6_{\pm 3.7}$
StereoViT-M	24.7	8.5	$\mathbf{56.7}_{\pm 3.4}$
StereoViT-L	**25.0**	**9.2**	$53.6_{\pm 3.4}$
GT depth	*26.8*	*15.2*	$55.0_{\pm 3.6}$

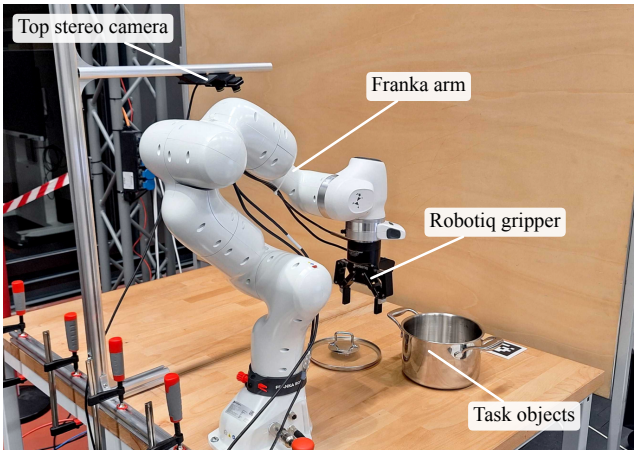


Figure S6. Real world robot arm setup.

Feature adapter. Features are average-pooled to the spatial grids in Tab. S6. The adapter applies per-token LayerNorm, a 1×1 convolution to 128 channels with GroupNorm and ReLU, three residual blocks with 3×3 convolutions and stride-2 downsampling, global average pooling, and an MLP (128 → 128 → 512). Tab. S6 summarizes the policy and training configurations. Both settings follow the respective official Diffusion Policy recipe.

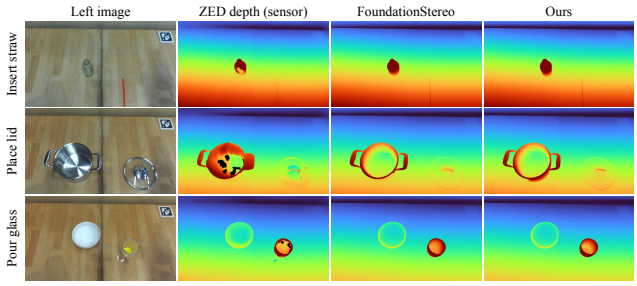


Figure S7. Depth on the real robot. The ZED sensor's depth has holes on metal and glass. FoundationStereo and our model recover these surfaces; ours also recovers the transparent pot lid and the beads inside the glass.

Baseline implementations. Our StereoPolicy [36] reimplementation combines a trainable ResNet-18 with FPN, frozen DINOv2 ViT-S/14 [75] features, and a two-layer stereo transformer with 2D RoPE, producing a 128-dimensional stereo feature. For frozen-feature baselines, we extract FoundationStereo's filtered cost volume following StereoVLA [21, 36], VGGT's final layer aggregation features following [74], and DAv2-B's four DPT-head features. These features use the same spatial pooling and adapter architecture, with the input projection adjusted to their channel dimension. DP3 [125] follows its official recipe with 1,024 farthest-point-sampled points.

Variation across training seeds. To assess the consistency of policy performance across training runs, we train three policies per task for StereoViT-M and the two baselines closest to it in Tab. 4. Each run has independent training seeds (0, 1, 2), and keeps the demonstrations, training configuration, and pretrained visual backbones fixed in Tab. S7. Each policy is evaluated on the same 100 episodes per task and setting as above, so all training runs and methods share scene seeds. We average success over the six tasks for each training run and report the mean and standard deviation over the three runs.

Table S6. **Diffusion Policy architecture and training configuration** for simulation (RoboTwin 2.0) and real-world robot experiments.

Configuration	RoboTwin 2.0	Real Robot
Pooled feature grid	15×20	18×32
U-Net channels	$[256, 512, 1024]$	$[512, 1024, 2048]$
Horizons (obs / pred / exec)	$3/8/6$	$2/16/8$
Robot	Dual-arm ALOHA-AgileX ($2\times$6-DoF + gripper)	7-DoF Franka arm + Robotiq 2F-85 gripper
Stereo camera	Head stereo camera, 240×320	ZED Mini (6.3 cm baseline), 1080×1920, 10 Hz
Proprioception	14D: 2 arms $\times$ (6 joints + gripper)	8D: 7 arm joints + gripper
Action space	14D joint targets: 2 arms $\times$ (6 joints + gripper)	10D EE target: pos. (3D) + 6D rot. + gripper
Image augmentation	None	Random 90% crop
Diffusion steps (train / inference)	100 DDPM / 100 DDPM	100 DDPM / 16 DDIM
Demonstrations per task	50	150
Optimizer	AdamW (lr $= 10^{-4}$), cosine schedule, batch size 128, 600 epochs, EMA	

Table S7. **RoboTwin 2.0 success rates across three training seeds** (%, 6-task average per run, and mean $\pm$ std across runs). Each run is evaluated on the same 100 episodes per task and setting on a single machine.

	Easy				Hard			
Method	S0	S1	S2	Mean	S0	S1	S2	Mean
StereoPolicy [36]	77.8	77.2	80.0	78.3±1.5	20.7	20.5	22.3	21.2±1.0
DAv2-B [119]	78.0	73.3	76.8	76.0±2.4	19.5	17.7	21.2	19.4±1.8
StereoViT-M	79.7	77.7	77.8	**78.4**±1.1	35.7	36.7	37.2	**36.5**±0.8

10. Real-World Experiments

Setup. A 7-DoF Franka arm with a Robotiq 2F-85 parallel-jaw gripper is observed by a ZED Mini stereo camera (6.3 cm baseline) mounted above the workspace and recorded at 10 Hz. The policy receives the rectified left/right pair. Demonstrations are collected by teleoperation in clean scenes. Evaluation uses unseen tablecloths and object instances (Fig. 5). The real-robot policies use Diffusion Policy's real-robot horizons and augmentation (Tab. S6).

Tasks. The six tasks of Tab. 5 and Fig. 5 are: *insert straw* (pick up a straw and insert it into a glass bottle), *place lid* (place a glass lid on a pot), *pour glass* (pick up a wine glass and pour its particles into a bowl), *place riser*, *hang cup*, and *stack cube*. In *place riser*, the robot picks up a 45 mm cube and places it on the top plate of a riser; risers range from 13 to 105 mm in height, with top plates of 100–140 mm. In *stack cube*, it stacks a 40 or 50 mm cube on a tower of one to three blocks, 25 to 105 mm high. Both tasks require judging the target's height: the 20 evaluation configurations form pairs that place the target at the same location but at two different heights, so that the two configurations look nearly identical in the image, and each task includes a height held out from the demonstrations (55 mm risers, 80 mm towers). In *hang cup*, the robot grasps a mug and hangs it on a mug stand; the positions of both and the direction of the mug's handle vary across configurations.

Data collection and evaluation. Each task has 150 teleoperated demonstrations, all recorded on the wooden training table with the objects placed at a different position and orientation in every demonstration. An AprilTag at the edge of the table serves to ground the table plane in the image and draw random object positions and orientations during data collection, ensuring a roughly equal distribution over the workspace. The object position and orientation to use for the next demonstration are shown to the operator on a nearby screen. We evaluate in unseen scenes: the wooden table is covered with a tablecloth never seen in training, and most tasks also swap or recolour objects. For *insert straw*, we use a blue reflective tablecloth and a green instead of a red straw; for *place lid*, a white tablecloth and a long-handled pan instead of the two-handled pot; for *pour glass*, a red and green checkered tablecloth and a purple bowl instead of the white bowl. For the three additional tasks, we change the evaluation setup as follows: for *place riser*, a white star-patterned tablecloth is used with the same cube and riser colors and shapes as in the training data, but adding more riser heights; for *hang cup*, a green tablecloth and a green instead of a blue mug; for *stack cube*, a red tablecloth with a green cube and a white block instead of the blue cube and yellow block used in demonstrations. Per task we fix 20 initial object placements, and rollout i starts from the same placement for every method. Consistency of the rollout starting placements is similarly ensured using the AprilTag. A rollout ends after 40 s or when the task is completed or aborted for hardware safety reasons, and counts as a success only if the whole task is completed.

Middlebury v3 (sorted by bad 4.0, all pixels)

Date	Name	Res	Weight Avg	Austr MP: 5.6 nd: 290 im0 im1 GT nonocc	AustrP MP: 5.6 nd: 290 im0 im1 GT nonocc	Bicyc2 MP: 5.6 nd: 250 im0 im1 GT nonocc	Class MP: 5.7 nd: 610 im0 im1 GT nonocc	ClassE MP: 5.7 nd: 610 im0 im1 GT nonocc	Compu MP: 1.5 nd: 256 im0 im1 GT nonocc	Crusa MP: 5.5 nd: 800 im0 im1 GT nonocc	CrusaP MP: 5.5 nd: 800 im0 im1 GT nonocc	Djemb MP: 5.7 nd: 320 im0 im1 GT nonocc	DjembL MP: 5.7 nd: 320 im0 im1 GT nonocc	Hoops MP: 5.7 nd: 410 im0 im1 GT nonocc	Livgrm MP: 5.9 nd: 320 im0 im1 GT nonocc	Nkuba MP: 5.5 nd: 570 im0 im1 GT nonocc	Plants MP: 5.6 nd: 320 im0 im1 GT nonocc	Stairs MP: 5.2 nd: 450 im0 im1 GT nonocc
11/04/26	StereoViT	F	**1.52** 1	2.24 6	**1.43** 1	0.97 2	2.90 10	5.40 18	0.76 2	1.46 5	**0.96** 1	0.86 3	**0.89** 1	2.63 2	1.04 3	**1.08** 1	1.83 5	0.34 6
11/08/26	SGT	F	1.59 2	2.14 5	1.78 7	0.97 2	0.64 2	9.89 66	2.08 15	1.32 3	**0.96** 1	**0.84** 1	0.96 2	**2.14** 1	**0.92** 1	1.13 2	**1.58** 1	0.28 3
06/27/25	S2M2	F	1.63 3	**1.62** 1	1.57 2	0.99 5	3.57 19	3.52 8	1.89 13	**1.07** 1	1.00 3	0.98 5	1.29 7	3.49 4	1.38 6	1.23 5	1.63 2	**0.25** 1
03/04/26	DepthFocus	F	1.67 4	1.82 3	1.69 6	1.03 6	3.13 12	3.24 6	1.39 8	1.24 2	1.14 4	**0.84** 1	1.22 5	5.93 10	1.06 4	1.34 6	1.72 4	0.32 5

Spring (sorted by 1px total)

	Name	1px ▲ total	1px low-detail	1px high-detail	1px matched	1px unmatched	1px not sky	1px sky	1px s0-10	1px s10-40	1px s40+	Abs	D1
1	StereoViT-L	5.813	5.566	20.708	4.916	23.622	5.492	10.692	3.808	5.616	10.203	0.400	2.206
2	StereoViT-M	6.067	5.807	21.723	5.085	25.560	6.114	5.351	3.126	6.226	11.169	0.428	2.410
3	RoCo-19-CorrDrop-Stereo	7.095	6.777	26.261	5.898	30.860	7.287	4.179	3.050	7.571	13.345	0.457	2.615
4	RoCo-22_StableBridge	7.135	6.824	25.908	5.940	30.861	7.371	3.554	2.936	7.757	13.244	0.471	2.705

KITTI 2015 (sorted by D1-all)

	Method	Setting	Code	D1-bg	D1-fg	D1-all	Density	Runtime	Environment	Compare
1	StereoViT-L			**0.95 %**	2.47 %	**1.20 %**	100.00 %	0.3 s	1 core @ 2.5 Ghz (C/C++)	
2	ForceStereo			1.09 %	**2.09 %**	1.26 %	100.00 %	0.13 s	Pytorch @ NVIDIA L20	
3	BC-Stereo			1.02 %	2.54 %	1.27 %	100.00 %	0.42 s	GPU @ 2.5 Ghz (Python)	
4	SGT			1.00 %	2.66 %	1.27 %	100.00 %	0.2 s	1 core @ 2.5 Ghz (Python)	

KITTI 2012 (sorted by Out-Noc, 3 px)

	Method	Setting	Code	Out-Noc	Out-All	Avg-Noc	Avg-All	Density	Runtime	Environment	Compare
1	StereoViT-L			**0.65 %**	**0.78 %**	0.3 px	0.3 px	100.00 %	0.3 s	1 core @ 2.5 Ghz (C/C++)	
2	SGT			0.68 %	0.85 %	**0.3 px**	0.3 px	100.00 %	0.2 s	1 core @ 2.5 Ghz (Python)	
3	sd_add			0.68 %	0.85 %	0.3 px	0.3 px	100.00 %	1 s	1 core @ 2.5 Ghz (C/C++)	
4	GeoAtt			0.70 %	0.88 %	0.3 px	0.4 px	100.00 %	0.1 s	1 core @ 2.5 Ghz (Python)	

ETH3D low-res two-view (test set, all pixels, sorted by bad 2.0)

Method	Info	all	lakes. 1l	lakes. 1s	sand box 1l	sand box 1s	stora. room 1l	stora. room 1s	stora. room 2l	stora. room 2s	stora. room 2 1l	stora. room 2 1s	stora. room 2 2l	stora. room 2 2s	stora. room 3l	stora. room 3s	tunnel 1l	tunnel 1s	tunne 2
StereoViT-L (ZS)		**0.04** 1	**0.03** 1	**0.15** 1	0.01 24	0.02 264	0.08 10	0.01 179	0.33 59	0.07 3	0.02 53	0.02 3	0.02 42	**0.00** 1	**0.03** 1	0.03 5	**0.00** 1	0.01 347	**0.0**
NBS		**0.04** 1	0.13 132	0.37 21	**0.00** 1	**0.00** 1	0.04 3	**0.00** 1	0.16 7	**0.06** 1	**0.00** 1	0.02 3	**0.00** 1	**0.00** 1	0.06 4	0.04 13	**0.00** 1	**0.00** 1	**0.0**
BC-Stereo	C	0.05 3	0.10 73	0.26 7	0.02 114	0.01 173	0.15 26	**0.00** 1	0.15 6	0.08 5	0.01 17	0.02 3	0.03 119	**0.00** 1	0.12 25	0.06 35	**0.00** 1	**0.00** 1	**0.0**

Figure S8. **Official leaderboards** screenshots; our anonymous entries are highlighted. All other metrics remain available on the benchmark websites. StereoViT-L is the single fine-tuned checkpoint of Sec. 3.2; its ETH3D entry is zero-shot.