STEREOVLM: BINOCULAR VISION-LANGUAGE MODELS FOR METRIC 3D SPATIAL REASONING

Anonymous authors
Paper under double-blind review

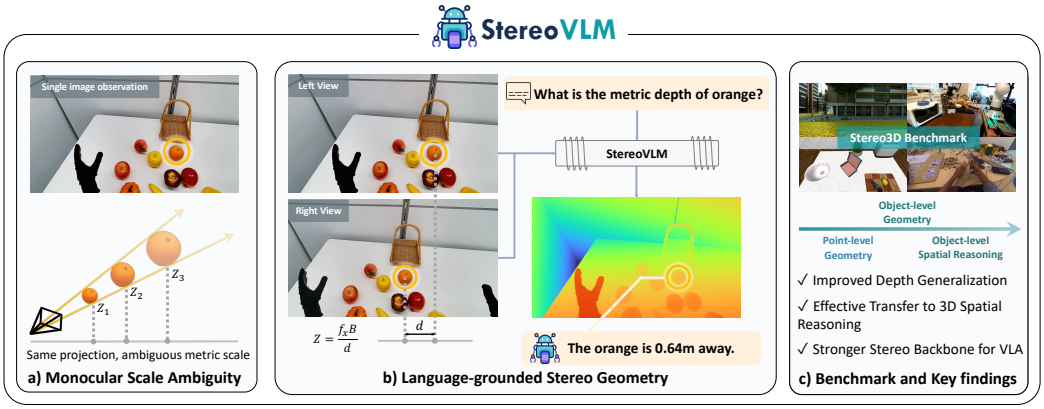

Figure 1: StereoVLM: from binocular geometry to spatial understanding and action. (a) A single image admits different object depths and sizes with the same projection, leaving metric scale ambiguous. (b) StereoVLM leverages binocular geometry from calibrated stereo observations to better answer language queries about metric depth. (c) Stereo3D-Bench spans point-level geometry, object-level geometry, and object-level spatial reasoning. Disparity supervision improves depth generalization, while binocular training transfers to broader 3D spatial reasoning and improves downstream VLA performance.

ABSTRACT

Vision-language models (VLMs) exhibit strong semantic understanding but remain limited in their ability to reason about metric 3D space in the physical world. Existing VLMs typically infer 3D scene structure from one or more images without known camera calibration or cross-view geometric constraints. This makes metric 3D inference inherently ambiguous and leaves it unclear whether their spatial predictions are grounded in visual evidence or driven primarily by semantic priors. In contrast, calibrated stereo provides explicit cross-view geometry that resolve this ambiguity, offer a more reliable foundation for metric 3D perception and geometric generalization. We investigate how VLMs can learn transferable 3D representations directly from calibrated binocular observations. To this end, we introduce **StereoVLM**, the first geometry-aware binocular VLM for metric 3D spatial reasoning. StereoVLM follows a hierarchical learning paradigm spanning from cross-view correspondence and disparity estimation to calibration-conditioned metric depth perception and semantic spatial reasoning. We further find that disparity supervision generalizes substantially better than direct binocular depth supervision on unseen environments. To evaluate these capabilities, we introduce **Stereo3D-Bench**, a comprehensive benchmark across three levels, from point-level geometry to object-level spatial reasoning. StereoVLM-8B achieves 95.12% depth accuracy (δ_1) and a 66.98 average object-level score on Stereo3D-Bench. Binocular training yields average gains of 3.54 points across five spatial benchmarks and 4.3 percentage points in RoboCasa manipulation success. Intervention results further indicate that these gains depend on valid cross-view geometry and cannot be explained solely by additional visual tokens or semantic priors.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in understanding and reasoning about visual content in the digital world (Alayrac et al., 2022). Building upon these advances, recent efforts have begun to extend such capabilities beyond the digital realm into the physical world, positioning VLMs as a key foundation for embodied AI (Driess et al., 2023; Kim et al., 2024; Brohan et al., 2023). However, extending VLMs from the digital to the physical world requires more than semantic understanding: it demands geometrically grounded spatial intelligence for reasoning and acting in 3D environments.

To extend VLMs beyond semantic understanding, recent studies have explored various ways to enhance their spatial reasoning capabilities. SpatialVLM (Chen et al., 2024) learns qualitative and quantitative spatial concepts from large-scale spatial question-answering data, while Spatial-RGPT (Cheng et al., 2024) augments region-level representations with monocular depth cues for object-centric reasoning. 3D-LLM (Hong et al., 2023), in contrast, grounds language models in reconstructed 3D scene representations. Despite these advances, spatial benchmarks such as VSI-Bench reveal persistent limitations in current models' ability to construct and reason about 3D space from visual observations (Yang et al., 2025). Moreover, the geometric information available to these models is typically inferred from monocular observations or supplied through estimated depth and reconstructed 3D representations, making it difficult to determine whether their spatial predictions are grounded in geometric evidence directly available from visual observations or primarily reflect learned priors—a concern reinforced by evidence-absent evaluations showing that VLMs can exploit scene- and category-level priors (Zhang et al., 2026).

Calibrated stereo provides explicit geometric constraints for metric 3D perception: cross-view correspondences determine disparity, which can be converted into metric depth using known camera intrinsics and baseline. Yet, providing binocular observations alone does not guarantee that a VLM will acquire geometric representations that generalize beyond the training distribution. Existing stereo methods primarily focus on geometric estimation, while VLMs typically treat paired images as generic visual inputs without explicitly learning the geometric relations between them. **This raises a central question:** *How can a VLM learn transferable representations from binocular geometry that support metric 3D perception and high-level spatial reasoning?* To investigate this question, we introduce StereoVLM, a geometry-aware binocular VLM designed to learn transferable 3D representations directly from calibrated stereo observations.

To connect binocular geometry with spatial understanding, we organize the training supervision into a hierarchy of complementary tasks spanning pixel-level correspondence, disparity estimation, calibration-conditioned metric depth, and object-level spatial reasoning. This hierarchy progressively connects cross-view matching to image-space geometric measurement and then to metric-scale estimation, before extending the learned geometry to language-referred objects and task-relevant 3D grounding. Importantly, we find that disparity provides a substantially more generalizable supervision signal than direct binocular depth on unseen environments, motivating its role as an intermediate geometric representation between cross-view correspondence and metric depth. We further establish Stereo3D-Bench to evaluate this progression across point-level geometry, object-level geometry, and object-level spatial reasoning. Beyond conventional task accuracy, controlled right-view interventions and geometric out-of-distribution evaluation test whether model predictions genuinely depend on valid cross-view geometry. The resulting binocular representations further transfer beyond stereo-specific tasks, improving the average score across five external spatial-reasoning benchmarks by 3.54 points and downstream RoboCasa manipulation success by 4.3 percentage points over matched training without stereo instructions.

Our contributions are:

- We introduce StereoVLM, the first geometry-aware binocular VLM for metric 3D spatial reasoning, with hierarchical supervision that connects cross-view correspondence and disparity to calibration-conditioned metric depth and language-grounded spatial reasoning, yielding transferable 3D spatial representations.
- We establish Stereo3D-Bench, a comprehensive benchmark spanning point-level geometry, object-level geometry, and object-level spatial reasoning, together with a verifiable evaluation protocol that uses controlled input interventions and geometric out-of-distribution evaluation to test whether predictions genuinely rely on binocular geometry.

- We demonstrate that disparity provides a more generalizable geometric supervision signal than direct binocular depth on unseen environments, while the resulting binocular representations transfer to broader spatial reasoning and downstream embodied control.

2 Related Work

Spatial reasoning in vision-language models Existing studies have enhanced spatial reasoning in VLMs through spatial instruction tuning (Chen et al., 2024), depth-aware visual representations (Cheng et al., 2024), and 3D grounding (Hong et al., 2023; Wang et al., 2025b). Meanwhile, benchmarks such as VSI-Bench (Yang et al., 2025) reveal that current multimodal models still struggle with metric 3D understanding and spatial reasoning. Recent works have further explored metric depth perception in VLMs at different granularities, with DepthLM (Yeh et al., 2026) and VLM3 (Cai et al., 2026) studying queried-point metric depth prediction, and DepthVLM (Yu et al., 2026) focusing on dense per-pixel metric depth estimation. However, these methods primarily focus on monocular, point- or pixel-level geometric perception, leaving its connection to higher-level semantic spatial reasoning and the use of cross-view geometric evidence underexplored.

Stereo vision and embodied intelligence Stereo vision recovers scene geometry by establishing cross-view correspondence and converting disparity into depth using camera calibration. Modern stereo methods, including RAFT-Stereo (Lipson et al., 2021), IGEV-Stereo (Xu et al., 2023), FoundationStereo (Wen et al., 2025) and S2M2 (Min et al., 2025), employ correlation or geometry volumes to achieve accurate and generalizable disparity estimation. Recent work has begun incorporating binocular sensing into embodied systems. StereoVLA (Deng et al., 2025) fuses stereo-geometric and monocular-semantic features and adds auxiliary depth and camera-parameter objectives for VLA manipulation, while StereoPolicy (Han et al., 2026) uses cross-view attention to adapt diffusion policies and pretrained VLA policies without explicit 3D reconstruction. EATR-Stereo (Wu et al., 2026) further routes paired-view evidence according to proprioceptive state for long-horizon humanoid control, and VLBiMan (Zhou & Jia, 2026) combines vision-language grounding with calibrated stereo reconstruction in a modular one-shot bimanual system. These methods demonstrate the value of binocular cues, with a primary focus on geometric estimation or embodied action. Our work bridges embodied stereo sensing and VLM spatial reasoning through a stereo VLM jointly trained on cross-view correspondence, disparity estimation, calibration-conditioned metric depth perception, and high-level semantic spatial reasoning.

3 Stereo3D-Bench

Stereo3D-Bench evaluates whether binocular evidence supporting metric geometry also transfers to language-grounded understanding of objects and task-relevant 3D structure. As illustrated in fig. 2, the benchmark organizes its tasks into three levels: point-level geometry, object-level geometry, and object-level spatial reasoning. The stereo pairs are drawn from four data sources spanning real-world and simulated environments. Controlled right-view interventions further test whether predictions depend on valid cross-view correspondence rather than merely on the presence of a second image. The supplementary material details data and question construction (section A), split and leakage controls (section B), and the evaluation protocol (section C).

3.1 Task Hierarchy

Stereo3D-Bench comprises eight tasks across three levels: three point-level geometry tasks, three object-level geometry tasks, and two object-level spatial reasoning tasks. Each query is posed as a short language question on an ordered stereo pair. Depending on the task, the model returns a coordinate, a scalar, a 2D bounding box, or an oriented 3D bounding box with an object label.

Point-level geometry. Given a queried pixel in the left image, the model predicts its corresponding location in the right image, its horizontal disparity, or its metric depth along the left-camera optical axis. These tasks assess the model's ability to establish cross-view correspondence, estimate binocular disparity, and recover metric scene geometry from stereo observations.

Object-level geometry. The same three geometric tasks are applied to an object specified by language. Object correspondence returns its matching bounding box in the right image, object disparity

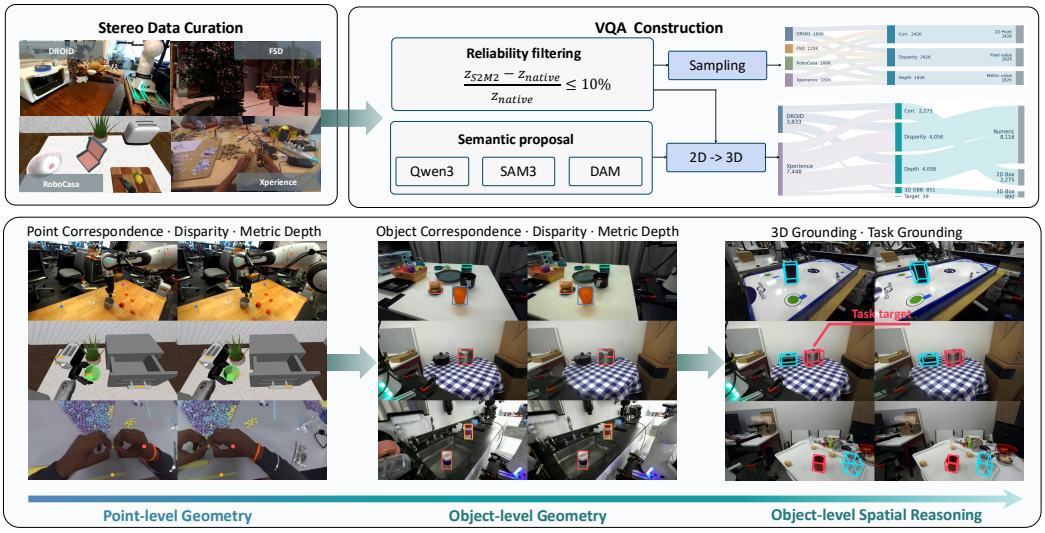

Figure 2: Overview of **Stereo3D-Bench**. Top: Real-world and simulated stereo data sources, with task and answer-format distributions for point- and object-level queries. Bottom: Examples of eight tasks organized into three levels: point-level geometry (correspondence, disparity, and metric depth), object-level geometry (object correspondence, disparity, and depth), and object-level spatial reasoning (oriented 3D grounding and task-target grounding).

predicts its representative disparity, and object depth predicts its representative metric depth along the left-camera optical axis.

Object-level spatial reasoning. Oriented 3D grounding identifies a referred object and predicts its label and oriented 3D bounding box in the rectified-left camera frame. Task-target grounding instead receives a manipulation instruction, identifies the object to be manipulated, and predicts that object's oriented 3D bounding box.

Together, these tasks test whether a model can progress from matching individual pixels, to associating geometry with named objects, and to grounding task-relevant objects in metric 3D space.

3.2 Binocular Controls

A correct prediction does not by itself demonstrate that the model uses binocular evidence. We evaluate each query under five input conditions while holding the checkpoint, question, calibration, and decoding procedure fixed. *Correct stereo* $(\mathbf{I}_l, \mathbf{I}_r)$ defines the primary benchmark score; the other four conditions are diagnostic controls. *Single left* supplies only $\mathbf{I}_l$, whereas *zero right* supplies $(\mathbf{I}_l, \mathbf{0})$, where $\mathbf{0}$ is a zero-valued image with the same dimensions as $\mathbf{I}_r$. These controls test whether any improvement requires visual evidence from the right image rather than merely a second image slot. *Duplicated left* supplies $(\mathbf{I}_l, \mathbf{I}_l)$, removing the binocular displacement while preserving the same scene content. *Shuffled right* supplies $(\mathbf{I}_l, \mathbf{I}'_r)$, where $\mathbf{I}'_r$ is taken from another benchmark sample. It preserves a natural second image but breaks the cross-view correspondence. Together, these controls test whether predictions depend on the geometrically matching right view.

4 Method

StereoVLM learns binocular spatial representations through hierarchical supervision spanning correspondence, disparity, calibration-conditioned metric depth, and language-grounded 3D reasoning. We introduce the stereo geometry followed by point- and object-level supervision.

4.1 Stereo Geometry Preliminaries

Consider a point observed at $\mathbf{p}_l = (u_l, v_l)$ in the left image and at $\mathbf{p}_r = (u_r, v_r)$ in the right image, where u and v denote horizontal and vertical pixel coordinates, respectively. For a rectified stereo pair, corresponding points lie on the same epipolar row up to residual calibration and discretization

errors, so $v_l \simeq v_r$. We define the horizontal disparity as

$$d(\mathbf{p}_l) = u_l - u_r. \tag{1}$$

Given the horizontal focal length f_x and stereo baseline B, disparity determines the point's metric depth along the left-camera optical axis:

$$Z(\mathbf{p}_l) = \frac{f_x B}{d(\mathbf{p}_l)}, \qquad d(\mathbf{p}_l) > 0. \tag{2}$$

Both d and f_x are measured in pixels. Thus, when B is expressed in meters, Z is obtained in meters.

The corresponding 3D point in the left-camera coordinate frame is obtained by back-projecting the pixel with the left-camera intrinsic matrix $\mathbf{K}_l$:

$$\mathbf{P}_l = Z(\mathbf{p}_l)\, \mathbf{K}_l^{-1} \begin{bmatrix} u_l & v_l & 1 \end{bmatrix}^\top. \tag{3}$$

These relations connect cross-view correspondence, image-space disparity, metric depth, and camera-frame 3D position.

4.2 Geometry-Consistent Stereo Supervision

We construct three point-level question types—correspondence, disparity, and metric depth—as complementary supervision targets derived from the same underlying stereo geometry. Given a queried left-image point $\mathbf{p}_l$, correspondence predicts its matching location $\mathbf{p}_r$ in the right image. Disparity predicts their horizontal displacement as defined in eq. (1), while metric depth along the left-camera optical axis is obtained from disparity and camera calibration according to eq. (2).

The three question types supervise different aspects of the same binocular relation. Correspondence teaches the model where a point matches across views, disparity quantifies the resulting image-space displacement, and depth associates that displacement with physical scale. For each metric-depth question, we provide the applicable camera calibration in the query and compute the target from the corresponding disparity using that sample's f_x and B. Thus, depth is not an independent geometric label; it is a calibration-conditioned expression of the underlying binocular displacement.

These questions are interleaved during instruction tuning, encouraging the model to represent cross-view matching, image-space displacement, and metric depth consistently.

4.3 Language-Grounded 3D Supervision

Point-level supervision captures binocular geometry at queried image locations, but does not associate that geometry with semantic entities. We therefore extend the same geometric relations to objects specified by language. Object correspondence predicts the bounding box of a referred object in the right image, while object disparity and depth predict the horizontal disparity and metric depth of its center point, respectively. These tasks connect correspondence, disparity, and depth with object identity and spatial extent, rather than with an isolated pixel alone.

We further supervise oriented 3D box grounding, which predicts an object's center, dimensions, and orientation in the rectified-left camera frame. This extends object-level depth from a single distance measurement to a complete spatial representation expressed in the same rectified-left camera frame as eq. (3). Finally, task-target grounding takes a manipulation instruction, identifies the object required by that instruction, and predicts its oriented 3D bounding box.

Together, the point- and object-level tasks form a hierarchy from binocular matching and geometric measurement to language-grounded object localization and task-relevant 3D understanding.

5 Experiments

5.1 Experimental Setup

We provide two variants, StereoVLM-2B and StereoVLM-8B, initialized from the corresponding Qwen3-VL-Instruct (Bai et al., 2025) checkpoints. Both models are instruction-tuned for 3,000 steps with all parameters updated, using an initial learning rate of 1×10^{-5}, a warmup ratio of 0.03, and an effective global batch size of 64. All stereo images have a resolution of 800×320.

Table 1: Geometry evaluation on Stereo3D-Bench. † denotes depth recovered from disparity.

Training Supervision	Standard Domains			Unseen		
	$\delta_1 \uparrow$	MAE ↓	AbsRel ↓	$\delta_1 \uparrow$	MAE ↓	AbsRel ↓
Monocular depth	77.91	14.08	18.48	51.00	26.08	36.75
Binocular depth	87.26	10.09	10.19	45.89	24.64	33.69
Binocular disparity†	**91.07**	**8.68**	**9.60**	**74.92**	**18.65**	**17.88**

Table 2: Comparison of direct and disparity-derived depth predictions from the same jointly trained models.

	StereoVLM-2B		StereoVLM-8B	
	Direct	Disparity	Direct	Disparity
MAE ↓	**7.52**	7.57	**6.89**	7.06
RMSE ↓	**32.54**	32.68	**29.93**	32.63
AbsRel ↓	**6.99**	7.44	**6.55**	6.97
$\delta_1 \uparrow$	**94.51**	94.46	**95.17**	95.15

Table 3: Effect of adding point-level geometry supervision to object-level training in StereoVLM-2B.

Supervision	Point-Level Geometry			Object-Level Reasoning	
	Corr. MAE ↓	Disp. MAE ↓	Depth $\delta_1 \uparrow$	3D IoU ↑	Joint@0.25 ↑
Object-level reasoning	59.98	22.65	0.43	24.64	56.75
+ Point-level geometry	**7.17**	**7.14**	**86.38**	**24.79**	**61.51**

Training combines binocular instructions with general visual-language, grounding, spatial-reasoning, tracking, and robotics data. Binocular instructions constitute 25% of the sampled mixture, comprising 15% point-level geometry and 10% object-level semantics. The source-level inventory, data expansion, sampling weights, and complete optimization details are provided in section D.1.

We strictly separate training and test data across all experiments and use a fully held-out in-house benchmark to assess generalization. We evaluate StereoVLM on Stereo3D-Bench, external spatial-reasoning benchmarks, and RoboCasa under their respective protocols. Experiment-specific settings are detailed in the corresponding subsections.

5.2 Learning Transferable Binocular Representations

We analyze how the choice of geometric supervision affects metric-depth generalization. We compare monocular depth, binocular depth, and binocular disparity under a controlled geometry-only setting. We then analyze whether joint disparity–depth training learns the calibrated conversion between the two quantities and whether geometric supervision benefits object-level semantic learning.

Which geometric target generalizes? We train three geometry-only variants using the same 234K rectified stereo pairs and optimization budget. The monocular-depth variant receives only the left image and directly predicts metric depth. The binocular-depth variant receives the ordered stereo pair and camera calibration and also predicts depth directly. The binocular-disparity variant receives the ordered pair and predicts horizontal disparity; at evaluation, its depth is recovered using the calibration through eq. (2). The training pool for these geometry-only variants contains Xperience-10M, DROID, RoboCasa, and an in-house robot-training collection. The final StereoVLM stereo-representation training mixture excludes RoboCasa, as detailed in section D.1. We evaluate both on the standard domains and on a held-out in-house benchmark whose scenes, images, and annotations are excluded from training.

Table 1 shows that binocular input alone does not guarantee better generalization. Direct binocular-depth supervision improves depth δ_1 over monocular supervision on the standard domains (87.26% vs. 77.91%), but performs worse on unseen scenes (45.89% vs. 51.00%). In contrast, binocular disparity supervision achieves the best results across all three depth metrics in both settings. On unseen scenes, its calibrated depth estimates reach 74.92% δ_1, exceeding direct binocular depth by 29.03 percentage points, with an MAE of 18.65 cm and an AbsRel of 17.88%. These results motivate disparity as an intermediate geometric representation for transferable metric 3D perception, rather than relying solely on direct depth supervision.

Consistency of direct and disparity-derived depth. We next examine whether joint disparity–depth supervision connects the two output spaces. Using the same jointly trained checkpoint and query points, we obtain depth through two routes: direct prediction from a metric-depth question, and calibrated conversion of the model's predicted disparity using eq. (2).

Table 4: Results on Stereo3D-Bench. General-purpose VLMs and StereoVLM receive the correctly ordered stereo pair, whereas monocular depth VLMs receive only the rectified left image. Correspondence and disparity MAE are measured in pixels; all other results are percentages. "–" denotes an unsupported task. Invalid outputs count as failures for δ_1 and the semantic metrics. Correspondence and disparity MAE exclude invalid predictions. Object-level Avg. is the unweighted mean of the five object-level tasks.

Model	Point-Level Geometry			Object-Level Tasks					
	Corr. MAE ↓	Disp. MAE ↓	Depth δ_1 ↑	Avg. ↑	Corr. mIoU ↑	Disp. Acc@2px ↑	Depth δ_1 ↑	3D Box mIoU ↑	Target Joint@0.25 ↑
General-Purpose VLMs									
Qwen3-VL-2B (Bai et al., 2025)	73.37	66.92	16.11	7.17	22.89	0.62	12.32	0.00	0.00
Qwen3-VL-8B (Bai et al., 2025)	46.87	41.25	21.47	9.12	24.14	1.23	20.11	0.12	0.00
GPT-5.6 Sol (OpenAI, 2026a)	28.03	26.36	17.67	9.54	27.83	7.79	11.75	0.34	0.00
GPT-6 Astra (OpenAI, 2026b)	9.27	7.55	75.96	47.61	68.88	45.05	93.40	17.92	12.82
Monocular Depth VLMs									
DepthLM-13B (Yeh et al., 2026)	–	–	28.40	–	–	–	–	–	–
DepthVLM-4B (Yu et al., 2026)	–	–	37.11	14.35	20.07	0.69	50.44	0.53	0.00
Binocular VLMs									
StereoVLM-2B	4.09	4.12	94.51	64.43	86.06	53.92	97.39	36.06	48.72
StereoVLM-8B	**3.80**	**3.76**	**95.12**	**66.98**	**87.62**	**60.52**	**97.76**	**37.70**	**51.28**

The two routes yield closely matched MAE and δ_1 at both model scales. Their δ_1 scores differ by only 0.05 percentage points for StereoVLM-2B and 0.02 points for StereoVLM-8B, while their MAEs differ by 0.05 cm and 0.17 cm, respectively. Direct prediction performs slightly better overall, with a larger advantage in 8B RMSE, suggesting that disparity conversion is more sensitive to occasional large errors. Overall, the comparable accuracy of the two routes supports that joint supervision connects predicted disparity with calibration-conditioned metric depth while retaining an effective direct-depth output.

Does semantic supervision replace geometric supervision? We replace all point-level geometry instructions with object-level instructions in StereoVLM-2B while keeping the total stereo instruction budget fixed. Point-level results are macro-averaged over the four shared evaluation domains, and object-level results are averaged over DROID and Xperience-10M. As shown in table 3, removing geometric supervision increases correspondence and disparity MAE from 7.17 to 59.98 pixels and from 7.14 to 22.65 pixels, respectively, while reducing depth δ_1 from 86.38% to 0.43%. Despite receiving more object-level instructions, the model's task-target success drops from 61.51% to 56.75%, whereas 3D-box mIoU remains similar (24.79% vs. 24.64%). These results indicate that additional object-level supervision cannot substitute for explicit point-level geometry supervision. Retaining geometric supervision preserves accurate point-level predictions and improves task-relevant 3D grounding under the same stereo instruction budget.

5.3 Main Results

We evaluate StereoVLM on the primary Stereo3D-Bench setting using the correctly ordered stereo pair. Our comparison includes two groups of baselines. General-purpose multi-image VLMs receive the same stereo pair and are evaluated on all tasks, allowing us to assess their binocular geometric capability. Monocular depth-specialized VLMs receive only the rectified left image and serve as references for single-view depth and, where supported, object-level prediction. Each baseline follows its official preprocessing and inference configuration, including its model-specific input resolution; predictions are mapped back to the benchmark coordinate system before evaluation. Results are reported only for tasks supported by each model's output interface. A dash denotes an unsupported task. Invalid outputs count as failures for δ_1 and the semantic metrics. Correspondence and disparity MAE are computed only over valid predictions.

Point-level geometry. StereoVLM-8B achieves the best results across all three point-level tasks. GPT-6 Astra, the strongest general-purpose baseline, shows substantial binocular capability, reaching correspondence and disparity errors of 9.27 and 7.55 pixels and a depth δ_1 of 75.96%. Its depth score falls to 43.88% under the single-view protocol in table 12, which removes both the right image and stereo baseline. StereoVLM-8B further reduces its correspondence and disparity errors by 5.47 and 3.79 pixels, respectively, and improves depth δ_1 by 19.16 percentage points to 95.12%. It also

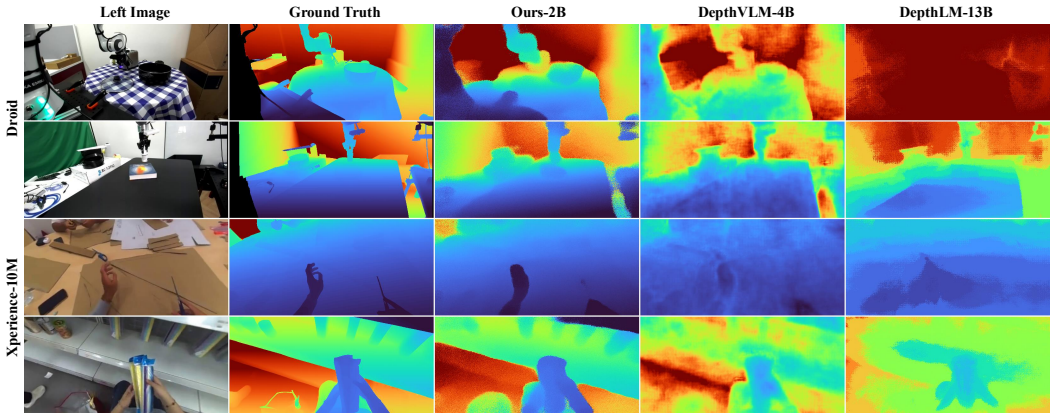

Figure 3: Qualitative comparison of metric-depth predictions on Stereo3D-Bench.

exceeds DepthVLM-4B, the strongest monocular depth baseline under its native single-view protocol, by 58.01 points in depth δ_1. The other evaluated general-purpose VLMs perform substantially worse on point-level geometry.

Language-grounded 3D reasoning. StereoVLM's gains extend from point-level geometry to language-referred objects and task-relevant 3D grounding. StereoVLM-2B and StereoVLM-8B achieve object-level averages of 64.43% and 66.98%, respectively, exceeding GPT-6 Astra's 47.61% by 16.82 and 19.37 percentage points. Relative to the single-view DepthVLM-4B reference at 14.35%, their gains are 50.08 and 52.63 points. On oriented 3D box and task-target grounding, StereoVLM-8B reaches 37.70% mIoU and 51.28% Joint@0.25, compared with 17.92% and 12.82% for GPT-6 Astra. GPT-6 Astra's object-depth δ_1 likewise falls from 93.40% with stereo input to 37.37% under the single-view protocol. Scaling from 2B to 8B yields consistent gains across all five object-level tasks.

Depth visualization. As shown in fig. 3, we qualitatively compare StereoVLM-2B with existing depth-specialized VLMs on representative samples from Stereo3D-Bench. StereoVLM-2B produces the most faithful depth reconstructions among the compared methods, recovering the overall scene geometry while better preserving object boundaries and foreground structure. Notably, StereoVLM-2B achieves these results with substantially fewer parameters than the two baselines: approximately half as many as DepthVLM-4B and one-sixth as many as DepthLM-13B.

Binocular evidence and single-view capability. We test whether StereoVLM's gains depend on a geometrically matched right view while also measuring the capability retained with only one image. Using StereoVLM-2B, we evaluate the five right-view input conditions defined in section 3.2. The checkpoint, left image, query, calibration, and decoding procedure are fixed, and only the right-view input is changed. In particular, the *single-left* condition retains the original stereo calibration, distinguishing this controlled intervention from the native single-view protocol in section D.2. Table 5 reports disparity, depth, and the average over the five object-level tasks.

Table 5: Effect of right-view interventions on StereoVLM-2B. Avg. denotes the mean across five object-level tasks.

Intervention	Geometry		Object-Level
	Disp. MAE ↓	Depth δ_1 ↑	Avg. ↑
Correct stereo	**4.09**	**94.51**	**64.43**
Duplicated left	32.85	7.40	8.18
Single left	9.07	80.82	40.26
Zero-right	9.31	79.70	39.81
Shuffled right	12.22	70.16	32.49

Dependence on matched right-view evidence. Correct stereo performs best across all reported metrics. Relative to single-left input, it reduces disparity MAE from 9.07 to 4.11 pixels, increases depth δ_1 from 80.82% to 94.96%, and raises the object-level average from 40.26% to 65.55%. Zero-right remains close to single-left, showing that preserving a two-image input structure without right-view content does not reproduce the stereo gain. Geometrically mismatched right views are harmful: shuffled-right reduces depth δ_1 to 70.16% and the object-level average to 32.49%. Duplicated-left causes the largest degradation, consistent with the model's sensitivity to the loss of expected cross-view displacement. Together, these interventions show that StereoVLM benefits from a geometrically matched right view rather than from an additional image slot or arbitrary second image.

Table 6: Effect of stereo instruction tuning on external spatial-reasoning benchmarks. Parentheses report absolute changes from training without stereo instructions.

	Single-Frame			Multi-Frame		Overall
Stereo	Omni3D	DA-2K	OmniSpatial	MindCube	VSI-Bench	Average
w/o	23.95	73.79	40.18	44.86	56.94	47.94
+stereo	**32.16**(+8.21)	**76.16**(+2.37)	**42.66**(+2.48)	**48.57**(+3.71)	**57.86**(+0.92)	**51.48**(+3.54)

Table 7: VLA policy success rates on RoboCasa under stereo observations.

Configuration	Success
Qwen3-VL-2B	34.9
+general	38.1
StereoVLM-2B (+stereo)	**42.4**

Retained single-view capability. The substantial performance retained under single-left input shows that StereoVLM can still rely on monocular appearance and learned scene priors when cross-view evidence is unavailable. Its close agreement with zero-right further indicates that an empty second image provides little additional information. StereoVLM therefore retains meaningful single-image capability while achieving its strongest geometric and object-level performance with valid binocular evidence.

5.4 Transfer Beyond Stereo3D-Bench

We examine whether stereo representation learning transfers beyond the questions in Stereo3D-Bench. We compare matched 2B representation models trained with and without stereo instructions. Both use the same initialization, optimization schedule, and training budget. When stereo data are removed, the remaining instruction categories are renormalized while preserving their relative sampling proportions. We evaluate the resulting representations on external spatial-reasoning benchmarks(Omni3D, DA-2K, OmniSpatial, MindCube, VSI-Bench) and downstream embodied control.

General spatial reasoning. Table 6 reports results on three single-frame benchmarks—Omni3D (Brazil et al., 2023), DA-2K (Yang et al., 2024), and OmniSpatial (Jia et al., 2026)—and two multi-frame benchmarks, MindCube (Wang et al., 2025a) and VSI-Bench (Yang et al., 2025). Adding stereo instructions improves all five benchmarks, increasing the overall average from 47.94 to 51.48. The largest gain occurs on Omni3D, which improves by 8.21 points, while the multi-frame MindCube and VSI-Bench scores improve by 3.71 and 0.92 points. These gains indicate that stereo representation learning transfers beyond explicit correspondence, disparity, and depth questions to broader spatial understanding.

Embodied control. We further evaluate whether the learned representation benefits downstream robotic control. We initialize the same GR00T policy (Bjorck et al., 2025) architecture from representation models trained with and without stereo instructions. Each policy is trained for 60k steps using 50 demonstrations per task and evaluated on 24 RoboCasa (Nasiriany et al., 2024) tasks with 50 rollouts per task. For this experiment, all RoboCasa stereo examples are excluded from representation training.

Detailed input controls in section D.4 further show that replacing the valid right image with a duplicated left image reduces success from 42.4% to 26.3% for the same stereo-policy checkpoint. A separately trained single-left policy reaches 40.0%. These results show that stereo representation learning provides a stronger initialization for embodied control and that the resulting stereo policy is sensitive to valid paired observations.

6 Conclusion and Limitations

We present StereoVLM, a geometry-aware binocular vision-language model for metric 3D spatial reasoning, trained with hierarchical stereo supervision from cross-view correspondence and disparity to calibration-conditioned metric depth and semantic spatial reasoning. We show that disparity supervision provides substantially stronger metric-depth generalization to unseen environments than direct binocular depth supervision, while controlled interventions verify that StereoVLM benefits from valid cross-view geometry. The resulting binocular representations further transfer to broader spatial reasoning and downstream embodied control. The current study is limited by its focus on controlled benchmarks and simulated kitchen tasks, as well as its reliance on calibrated and synchronized cameras; future work should address the simulation-to-real gap and improve robustness to challenging visual conditions and diverse robot platforms.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing, LaTeX editing, and figure prototyping. As part of the annotation pipeline, Qwen3 proposes candidate objects and DAM generates region descriptions for semantic consistency checks, as detailed in section A.3. Benchmark annotations undergo automatic geometric and semantic filtering and manual review to remove low-quality questions; training annotations undergo automatic filtering without manual review. Numerical benchmark answers are derived from source-native geometric references rather than generated by language models. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A large benchmark and model for 3D object detection in the wild. In *CVPR*, 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Zhipeng Cai, Zhuang Liu, Yunyang Xiong, Zechun Liu, Vikas Chandra, and Yangyang Shi. Vlm3: Vision language models are native 3d learners. *arXiv preprint arXiv:2605.30561*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *ICLR*, 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *NeurIPS*, 2024.

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. *arXiv preprint arXiv:2512.21970*, 2025.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

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.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 2023.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *ICLR*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. In *ICCV*, 2025.

Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*. IEEE, 2021.

Junhong Min, Youngpil Jeon, Jimin Kim, and Minyong Choi. S 2 m 2: scalable stereo matching model for reliable depth estimation. In *ICCV*. IEEE, 2025.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

OpenAI. GPT-5.6 Sol. OpenAI API model documentation, 2026a. URL `https://developers.openai.com/api/docs/models/gpt-5.6-sol`. Accessed: 2026-09-24.

OpenAI. GPT-6 Astra. OpenAI API model documentation, 2026b. URL `https://developers.openai.com/api/docs/models/gpt-6-astra`. Accessed: 2026-09-24.

Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *ICLR*, 2025a.

Yuxin Wang, Lei Ke, Boqiang Zhang, Tianyuan Qu, Hanxun Yu, Zhenpeng Huang, Meng Yu, Dan Xu, and Dong Yu. N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. *arXiv preprint arXiv:2512.16561*, 2025b.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *CVPR*. IEEE, 2025.

Songwei Wu, Rui Zhao, Fan Yang, Zhongqiang Nie, Zhiduo Jiang, Wandong Sun, Yuwei Li, Yang Liu, and Hong Liu. Eatr-stereo: Embodiment-aware routing of paired stereo evidence for humanoid vision-language-action control. *arXiv preprint arXiv:2608.17453*, 2026.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*. IEEE, 2023.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*. IEEE, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *NeurIPS*, 2024.

Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory P Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, Yangyang Shi, et al. Depthlm: Metric depth from vision language models. In *ICLR*, 2026.

Hanxun Yu, Xuan Qu, Yuxin Wang, Jianke Zhu, and Lei Ke. Unlocking dense metric depth estimation in vlms. *arXiv preprint arXiv:2605.15876*, 2026.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning. *arXiv preprint arXiv:2604.24300*, 2026.

Huayi Zhou and Kui Jia. Vlbiman: Vision-language anchored one-shot demonstration enables generalizable bimanual robotic manipulation. In *ICLR*, 2026.

SUPPLEMENTARY MATERIAL

This supplementary material documents the construction and evaluation of Stereo3D-Bench, together with additional training details and experiments for StereoVLM. Training and benchmark pairs follow a shared processing and annotation procedure but are kept separate by the data split. The sections below describe this procedure, the split, the evaluation protocol, and supporting results.

▷ **section A**: Geometric references, semantic annotations, and query and answer schemas, with point-level and object-level examples.
▷ **section B**: The training mixture, Stereo3D-Bench, and the held-out in-house benchmark, including per-source pair and QA counts.
▷ **section C**: Metric definitions, invalid outputs, and the implementation of right-view interventions.
▷ **section D**: Training settings, per-domain and intervention results, joint disparity–depth supervision, calibration perturbations, and external benchmark and RoboCasa protocols.

A DATA CONSTRUCTION

Training data and Stereo3D-Bench share the same stereo representation, question schemas, and coordinate conventions, but intentionally use different geometric-reference pathways. Training supervision uses FoundationStereo predictions, whereas Stereo3D-Bench is scored against source-native geometric references. This separation prevents benchmark performance from directly measuring reproduction of the stereo estimator used to construct the training data. The following subsections describe the shared representation, the training and evaluation reference pathways, the semantic annotations, and question generation. Dataset splitting and leakage control are described in section B.

A.1 UNIFIED STEREO DATA REPRESENTATION

Every source sample is converted into an ordered, rectified stereo pair at a resolution of 800×320. We retain the rectified camera metadata, including the horizontal focal length f_x, stereo baseline B, and camera extrinsics when provided by the source. Disparity is expressed in rectified-left pixel coordinates, and physical depth is represented by the left-camera optical-axis distance in meters.

Training and evaluation pairs use the same representation but differ in reference provenance. FoundationStereo (Wen et al., 2025) provides dense disparity supervision for the training mixture. In Stereo3D-Bench, DROID uses its aligned sensor depth, Xperience-10M uses its officially released stereo-depth annotation, RoboCasa uses simulator geometry, and FSD uses its source-native disparity reference. Because FSD does not provide physical metric calibration, it contributes correspondence and disparity questions but not metric-depth questions. Figure 4 compares the source-native references with the outputs of S2M2 and FoundationStereo.

A.2 POINT-LEVEL GEOMETRIC REFERENCES

Training references. For training samples with source-native geometric references, we retain pixels whose native reference is valid and agrees with FoundationStereo. FoundationStereo then provides the numerical disparity target on these filtered regions. When physical calibration is available, disparity is converted to metric depth as

$$Z = \frac{f_x B}{d}. \tag{4}$$

Thus, the training target is provided by FoundationStereo, while source-native references are used to identify reliable training regions.

Benchmark references. Stereo3D-Bench uses source-native geometry rather than FoundationStereo outputs. For DROID, the numerical depth target is the aligned ZED sensor depth. For Xperience-10M, it is the officially released metric stereo-depth annotation. For RoboCasa, depth and disparity are obtained directly from the simulator. FSD supplies source-native disparity but no physical metric-depth reference and is therefore excluded from metric-depth evaluation. When a physical depth reference Z is available, its disparity and right-image correspondence are derived

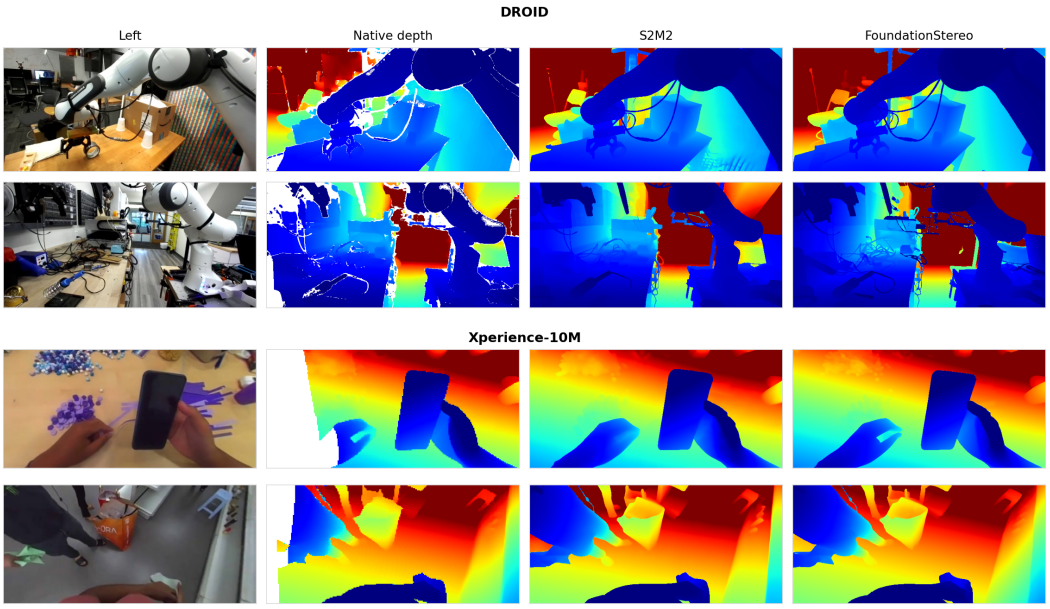

Figure 4: Comparison of source-native and estimated left-view depth. The first two rows show DROID examples and the last two show Xperience-10M examples. Columns show the rectified left image, the source-native depth, S2M2 depth, and FoundationStereo depth. Predicted depth is converted from disparity using the physical focal length and baseline released for each pair. All maps are aligned to the rectified-left image; colors are normalized independently within each row and should not be compared across rows. White denotes invalid depth in the corresponding map. Rectification borders are cropped for display.

using

$$d = \frac{f_x B}{Z}, \qquad x^r = x^l - d. \tag{5}$$

For the two real-image sources, DROID and Xperience-10M, we use S2M2 (Min et al., 2025), a separately trained stereo estimator that is not used to generate our training targets, as a reliability estimator. After converting its disparity to depth using the same physical calibration, a pixel is retained when both depths are valid and

$$\frac{|Z_{\text{S2M2}} - Z_{\text{native}}|}{Z_{\text{native}}} \leq 0.10. \tag{6}$$

S2M2 determines only the validity mask; all benchmark answers are still computed from the source-native reference. Importantly, FoundationStereo, which supplies the training supervision, is not used to select benchmark pixels or to define their answers.

We additionally remove pixels within 16 pixels of the image boundary and pixels whose projected right-image locations fall outside the common valid crop. Local neighborhood checks suppress isolated valid pixels and discontinuities unsupported by neighboring reference values. These checks are applied before any evaluated model is run, and sampling uses a fixed random seed.

Point sampling. Retained pixels are divided into a disparity-edge band and the remaining valid region. The edge band contains retained pixels within three pixels of a ridge in the source-native disparity map, ensuring that the candidate pool includes geometric boundaries rather than only large smooth surfaces. In the fixed Stereo3D-Bench protocol, each evaluated family contains 120 sampled points per pair, grouped into 15 questions of eight points each. Correspondence, disparity, and depth use separately sampled points but the same source-native reference and coordinate convention.

Figure 5 visualizes the reliability masks. The FoundationStereo agreement mask is included only as a diagnostic comparison with the training teacher; it is not used to construct Stereo3D-Bench.

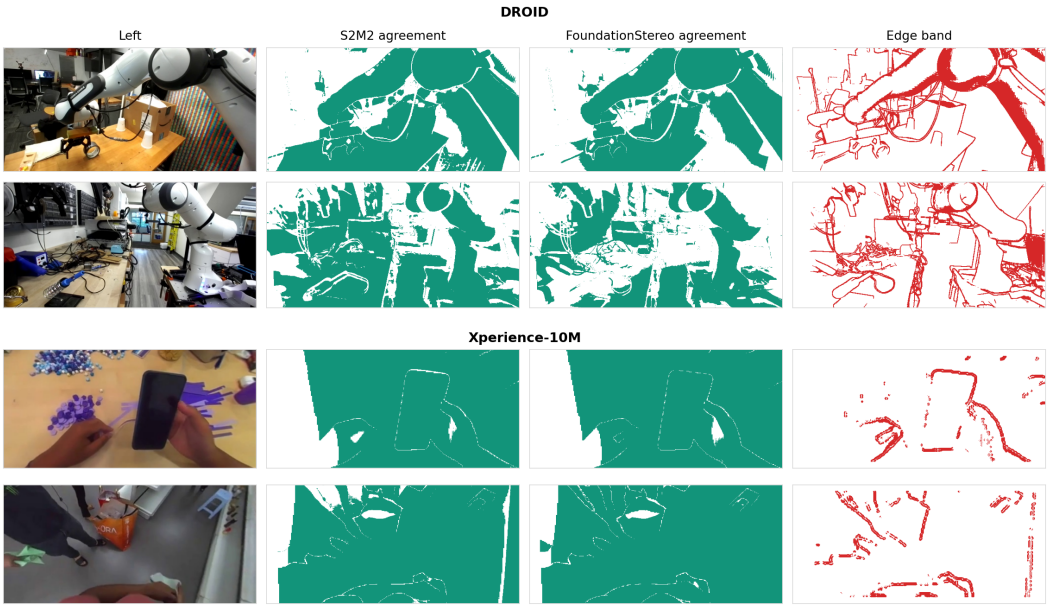

Figure 5: Point-level reference filtering. The first two rows show DROID and the last two show Xperience-10M. Green pixels in the second column have S2M2 depth within 10% of the source-native depth and form the benchmark validity mask. The third column shows the analogous FoundationStereo agreement mask for comparison only; it is not used for benchmark sampling. Red pixels in the final column lie within three pixels of a ridge in the source-native disparity map. Rectification borders are cropped for display.

A.3 Object-Level Semantic References

Object references are built from the rectified left image. Qwen3 (Bai et al., 2025) proposes object names, SAM3 (Carion et al., 2026) produces their masks, and DAM (Lian et al., 2025) supplies region descriptions used to reject semantically inconsistent proposals. The right image is not segmented. For benchmark pairs, object disparity, depth, right-image location, and lifted 3D geometry are computed from the source-native geometric references described in section A.2, rather than from FoundationStereo.

Valid object pixels are lifted into the left-camera point cloud and expressed in a gravity-aligned frame, where object extent and yaw are estimated. The resulting oriented 3D box is then transformed back to the rectified-left camera frame. DROID reaches the gravity-aligned frame using the provided camera extrinsics. For the egocentric Xperience-10M images, the gravity-aligned frame is estimated because the camera optical axis is not generally perpendicular to the ground.

These references instantiate the five object-level tasks defined in section 3.1: object correspondence, object disparity, object depth, oriented 3D grounding, and task-target grounding. fig. 6 illustrates the resulting references.

A.4 Question Generation and Quality Control

A question is formed from the references above and is posed on the ordered stereo pair. A point-level question lists one or more pixels in the left image, and the answer returns one value per pixel in the same order. Correspondence is a right-image coordinate, disparity is expressed on a normalized scale whose image width is 1000, and depth is in meters. The focal length and baseline are written into depth questions. An object-level question refers to an object by language. It asks for the object's right-image box, its disparity, its depth, its oriented 3D box in the left-camera frame, or the label and 3D box of the object to manipulate.

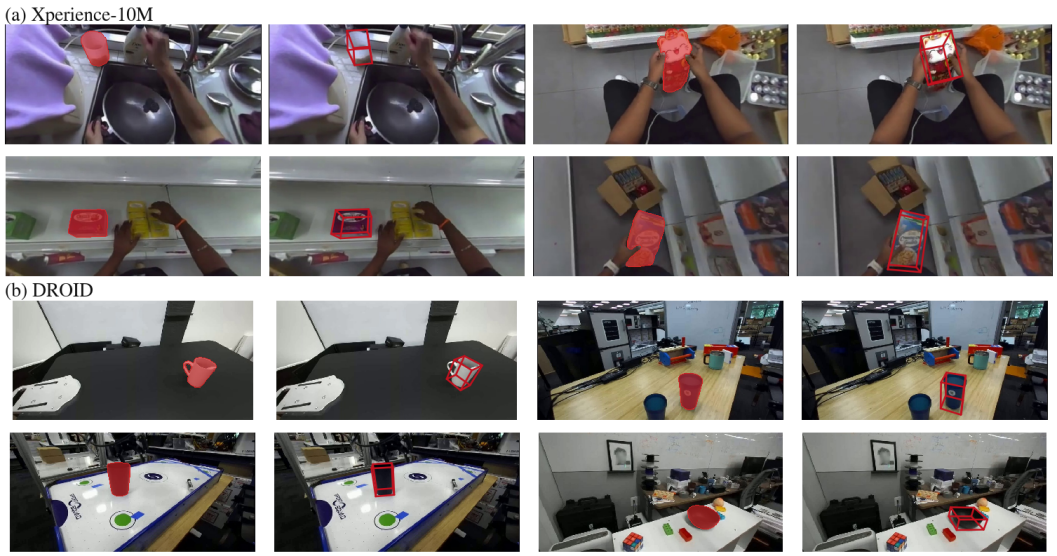

Figure 6: Object references. (a) Xperience-10M, using the estimated gravity-aligned frame. (b) DROID, using the provided camera extrinsics. In each pair, the left image shows the object mask and the right image projects the oriented box after it is written back to the left camera.

Both training and Stereo3D-Bench apply the same automatic filters. Part of the proposed object list is removed. A 3D box is kept only when the box projected back onto the image has sufficient overlap with the object mask. Stereo3D-Bench then receives a manual review, and questions that remain low quality are removed. Training questions do not receive this review.

A.5 Question and Answer Schemas

We generate questions from task-specific templates. Each task has a fixed answer schema and multiple question paraphrases: two each for point correspondence, disparity, and depth; three each for object correspondence, disparity, and depth; five for oriented 3D grounding; and three for task-target grounding. table 8 shows one paraphrase per task. Placeholders are replaced with the sample's coordinates, object reference, instruction, and calibration.

Point queries list one or more pixels in the left image. For multiple pixels, the answer contains one numbered line per pixel in the same order; a single-pixel query returns only the requested value. Disparities are reported in width-normalized coordinates, $d_{1000} = 1000 d_{\mathrm{px}}/W$, where W is the image width in pixels. Depth is the left-camera z coordinate in meters. Metric-depth questions state the rectified focal length and stereo baseline.

Object correspondence returns a right-image box $[x_1, y_1, x_2, y_2]$. Object disparity and depth return a single numeric value in the same units as their point-level counterparts. Oriented 3D grounding and task-target grounding return an object label and a 3D box expressed in the left-camera frame.

B Splits and Leakage Control

Training data and Stereo3D-Bench follow the data-construction procedure in section A, while their source data are separated before stereo pairs, geometric references, semantic annotations, or questions are generated. Consequently, no benchmark image or question derived from it is used during training. table 9 summarizes the resulting benchmark composition.

For DROID, splitting is performed at the scene level. We define a scene by the laboratory, robot, and collection date. All recordings, episodes, camera views, and frames associated with one scene are assigned to the same split. For Xperience-10M, the recording UUID is used as the scene-grouping

Table 8: Representative question templates. Angle-bracketed fields are replaced with sample-specific values.

Task	Question template
Point-level geometry	
Correspondence	For each of the following pixels in the left stereo image, locate the projection of the same point in the right stereo image. Pixels: <numbered pixels>. Answer one line per pixel as 1. (x, y), in the same order.
Disparity	For each of the following pixels in the left stereo image, give the horizontal disparity to its corresponding pixel in the right stereo image. Express each disparity in the normalized scale where the image width equals 1000. Pixels: <numbered pixels>. Answer one line per pixel as 1. <value>, in the same order.
Depth	For each of the following pixels in the left stereo image, estimate its depth along the camera z-axis using rectified stereo geometry. The stereo pair is rectified. The focal length is <fx> pixels and the baseline is <B> meters. Pixels: <numbered pixels>. Answer one line per pixel as 1. <value>, using meters, in the same order.
Object-level semantics	
Correspondence	In the left stereo image, identify <object>. Then locate the same physical object in the right stereo image and return its bounding box. Answer with the right-image bounding box only.
Disparity	In the left stereo image, identify <object>. What is its horizontal disparity in the right stereo image? Express the disparity in the normalized scale where the image width equals 1000. Answer with the numeric value only.
Depth	Stereo camera calibration for this displayed pair: fx=<fx> pixels, baseline=<B> meters. What is the z-depth of <object> in the left stereo image along the left camera z-axis? Answer with the numeric depth value in meters only.
3D grounding	Rectified-left camera calibration: fx=<fx>, fy=<fy> pixels; baseline=<B> meters. The principal point is at the image center. Predict the 3D oriented bounding box of <3d_ref> in the left stereo image. Answer with the 3D bounding box and object label only.
Task target	Rectified-left camera calibration: fx=<fx>, fy=<fy> pixels; baseline=<B> meters. The principal point is at the image center. For the task below, identify the visible object that should be manipulated and predict its 3D oriented bounding box in the left stereo image. Answer with the 3D bounding box and target label only. Task: <instruction>.

unit, and all episodes and frames from one recording remain together. Therefore, benchmark scenes from either source contribute neither images nor derived questions to training.

The remaining sources use source-specific exclusion units. FSD benchmark images are excluded from training at the source-frame level. RoboCasa contains 24 official task identities, five of which are reserved for evaluation. Samples from these five tasks are excluded from the training mixture, making the corresponding evaluation split disjoint at the task-identity level. FSD supplies correspondence and disparity references but no physical metric-depth reference; metric depth is therefore evaluated on the other 1,517 geometry pairs. RoboCasa semantic questions are excluded from the primary semantic benchmark.

The held-out in-house benchmark is collected independently of all sources above. None of its scenes, images, or derived annotations appears in either the training mixture or Stereo3D-Bench. We reserve it as a fully held-out evaluation of generalization to a new collection.

In the fixed Stereo3D-Bench geometry protocol, each evaluated family uses 15 grouped questions per pair. DROID, Xperience-10M, and RoboCasa contain three geometry families, whereas FSD contains correspondence and disparity only. The geometry QA counts in table 9 refer to the standard stereo input and do not count repeated evaluations under right-view interventions.

C Evaluation Protocol

Point-level geometry. Coordinates and disparities expressed on the width-1000 normalized scale are converted to pixels at the displayed image resolution before scoring. In particular, $d_{\mathrm{px}} = W d_{1000}/1000$ for image width W. Let $\mathcal{P}$ contain all queried points for a task, and let $\mathcal{P}_{\mathrm{valid}}$ contain those with a successfully parsed, finite prediction. The primary correspondence error is the

Table 9: Stereo3D-Bench composition by source. Geometry and semantic pairs are selected independently and therefore should not be summed to obtain a unique-image count.

Source	Geometry		Semantic	
	Pairs	QA	Pairs	QA
DROID	517	23,265	556	3,833
Xperience-10M	500	22,500	1,000	7,448
FSD	500	15,000	–	–
RoboCasa	500	22,500	–	–
Total	**2,017**	**83,265**	**1,556**	**11,281**

horizontal error of the disparity implied by the predicted right-image coordinate:

$$\mathrm{MAE}_{\mathrm{corr}} = \frac{1}{|\mathcal{P}_{\mathrm{valid}}|} \sum_{i \in \mathcal{P}_{\mathrm{valid}}} \left| (x_i^l - \hat{x}_i^r) - d_i \right|. \tag{7}$$

Here x_i^l is the queried left-image coordinate, $\hat{x}_i^r$ is the predicted right-image coordinate, and d_i is the reference disparity, all in pixels. The two-dimensional endpoint error and vertical coordinate error are retained as diagnostic metrics but do not enter the primary correspondence score. Disparity MAE is

$$\mathrm{MAE}_d = \frac{1}{|\mathcal{P}_{\mathrm{valid}}|} \sum_{i \in \mathcal{P}_{\mathrm{valid}}} |\hat{d}_i - d_i|, \tag{8}$$

also in pixels. The valid-point set is determined separately for each task; therefore, the correspondence and disparity denominators need not match.

For metric depth, we report MAE in centimeters, AbsRel, and δ_1:

$$\mathrm{MAE}_Z = \frac{1}{|\mathcal{P}_{\mathrm{valid}}|} \sum_{i \in \mathcal{P}_{\mathrm{valid}}} |\hat{Z}_i - Z_i|, \tag{9}$$

$$\mathrm{AbsRel} = \frac{1}{|\mathcal{P}_{\mathrm{valid}}|} \sum_{i \in \mathcal{P}_{\mathrm{valid}}} \frac{|\hat{Z}_i - Z_i|}{Z_i}, \tag{10}$$

$$\delta_1 = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}_{\mathrm{valid}}} \mathbf{1}\left[\max\left(\frac{\hat{Z}_i}{Z_i}, \frac{Z_i}{\hat{Z}_i} \right) < 1.25 \right]. \tag{11}$$

A valid depth prediction must additionally be positive. AbsRel is multiplied by 100 when reported as a percentage. MAE and AbsRel are computed over valid predictions, whereas the δ_1 denominator contains all queried points; invalid depth predictions therefore count as failures under δ_1. Point-level metrics weight queried points equally rather than first averaging within each question.

Object-level semantics. Object correspondence is measured by the IoU between the predicted and reference right-image boxes, averaged over all correspondence questions. An invalid box contributes an IoU of zero. Object disparity uses $\mathrm{Acc@2px}$, the fraction of questions satisfying $|\hat{d}-d| \leq 2$ pixels after conversion from the width-1000 normalized scale. Object depth uses the δ_1 definition above. Oriented 3D grounding uses the volumetric IoU between the predicted and reference oriented boxes in the rectified-left camera frame, averaged over all questions, with invalid boxes assigned zero IoU. Task-target grounding is correct only when the predicted target label is correct and its oriented 3D-box IoU is at least 0.25. The semantic average is the unweighted mean of these five task scores after each is expressed as a percentage.

Invalid outputs. A prediction is invalid if it is missing, cannot be parsed into the required answer schema, contains non-finite values, gives a non-positive depth, or specifies an invalid 3D box. For a multi-point answer, parsed entries are scored individually and missing entries are marked invalid; entries beyond the requested number are ignored. Invalid semantic predictions receive zero IoU or count as failures for $\mathrm{Acc@2px}$, δ_1, and task-target grounding. Invalid point-level predictions are excluded from MAE and AbsRel but remain in the coverage denominator; invalid depth predictions also remain in the δ_1 denominator.

Table 10: Stereo instruction data used for training. Counts are rounded for readability. Xperience-10M is reported after its fivefold expansion. Geometry and semantic pairs are constructed independently.

Source	Geometry		Semantic	
	Pairs	QA	Pairs	QA
DROID	100.0K	8.40M	85.2K	710K
FSD	1.00M	84.0M	–	–
Xperience-10M (5×)	410.3K	34.5M	–	–
Total	**1.51M**	**126.9M**	**85.2K**	**710K**

Controlled right-view inputs. For each query, the checkpoint, left image, question, calibration, and decoding configuration are held fixed. Correct stereo (I_l, I_r) provides the primary benchmark score. Four diagnostic conditions alter the availability or validity of the right view. Single-left input removes the second image, (I_l). Duplicated-left input places the left image in both slots, (I_l, I_l). Zero-right input replaces the right view with an equally sized black image, $(I_l, \mathbf{0})$. Shuffled-right input replaces the correct right image with a right image from another benchmark example, (I_l, I'_r). Scores under these four conditions are reported separately and do not enter the primary Stereo3D-Bench score.

D Additional Results

This section provides the training-data composition and optimization settings omitted from the main paper, followed by additional analyses of the single-view capability retained after stereo training, robustness to calibration errors, and transfer to embodied control. We conclude with qualitative results across geometric and object-level tasks.

D.1 Training Data and Optimization

Stereo training data. We construct the stereo instruction data from DROID, FSD, and Xperience-10M. All stereo pairs are processed using the construction procedure in section A. We exclude RoboCasa from stereo-representation training so that the embodied-control evaluation remains disjoint from the representation-training data.

table 10 reports the training inventory after filtering and data expansion. Geometry and semantic pairs are counted separately because the two annotation pipelines retain different subsets of the source images. Geometry QA counts denote complete question–answer records rather than individual points within a multi-point question. Semantic counts cover the five object-level task families used in the final model and benchmark.

For Xperience-10M, each of the 82,061 underlying stereo pairs is included five times in the constructed training inventory, together with its associated geometry questions. We therefore count 410,305 training-pair records and 34,465,620 geometry QA records for this source. This fivefold expansion is part of the reported training-data volume; it is not an online sampling multiplier and does not imply 410,305 distinct stereo image pairs.

Instruction mixture. Stereo instructions account for 25% of the final instruction mixture. We allocate 15% to point-level stereo geometry and 10% to object-level stereo semantics, preserving a $60{:}40$ ratio within the stereo portion. The remaining 75% is distributed among the original monocular instruction categories in proportion to their previous weights. The resulting top-level sampling probabilities are shown in table 11. These probabilities specify the composition of the instruction stream after construction of the stereo inventory above.

Optimization. We initialize each StereoVLM variant from the corresponding Qwen3-VL-Instruct checkpoint and perform full-parameter instruction tuning, including the vision encoder, language model, and multimodal aligner. All variants use the same data mixture and optimization schedule. Training runs for 3,000 optimization steps with an initial learning rate of 1×10^{-5}, a warmup ratio

Table 11: Top-level sampling probabilities in the final instruction mixture.

Instruction category	**Sampling probability**
General visual-language instruction	37.50
Grounding	9.93
Spatial reasoning	17.09
Tracking	3.25
Robotics	7.23
Stereo geometry	15.00
Stereo semantics	10.00
Total	**100.00**

Table 12: Results on Stereo3D-Bench under single-view input. All models receive only the rectified left image, and the stereo baseline is removed from the question. For models with invalid depth outputs, Depth δ_1 is computed over all queries, treating invalid outputs as failures.

	Geometry	**Semantic**		
Model	Depth $\delta_1 \uparrow$	Depth $\delta_1 \uparrow$	3D Ground. mIoU $\uparrow$	Target Joint@0.25 $\uparrow$
General-Purpose VLMs				
Qwen3-VL-2B (Bai et al., 2025)	19.80	28.56	0.06	0.00
Qwen3-VL-8B (Bai et al., 2025)	12.77	13.97	0.20	0.00
GPT-5.6 Sol (OpenAI, 2026a)	17.25	29.62	0.18	0.00
GPT-6 Astra (OpenAI, 2026b)	43.88	37.37	7.02	7.69
Binocular VLMs				
StereoVLM-2B	**73.51**	**81.52**	**16.92**	25.62
StereoVLM-8B	63.22	53.33	6.82	**28.21**

of 0.03, and an effective global batch size of 64. We use a maximum sequence length of 32,768 tokens, bfloat16 computation, gradient checkpointing, and DeepSpeed ZeRO-2. Stereo images are presented at 800×320, the same resolution used to construct the geometric and semantic references. Dataset and training-order shuffling use a fixed random seed, and all reported model variants differ only in the explicitly stated experimental factor.

D.2 Single-View Capability and Input Controls

Although StereoVLM is trained with explicit stereo supervision, it is not restricted to binocular inputs. To evaluate the monocular capability retained by the model, we remove the right image and provide only the rectified left image. We also remove the stereo baseline from the question. The checkpoint, focal length, remaining question text, and decoding configuration are otherwise unchanged. This setting prevents the model from using binocular correspondence and measures what can be recovered from monocular appearance, object-scale priors, and the visual-language knowledge inherited from pretraining.

table 12 shows that StereoVLM retains substantial single-view depth and spatial reasoning capability. StereoVLM-2B and StereoVLM-8B obtain point-level depth δ_1 scores of 73.51 and 63.22, respectively, compared with 43.88 for the strongest general-purpose baseline. StereoVLM-2B also reaches an object-depth δ_1 of 81.52. For oriented 3D grounding, StereoVLM-2B and StereoVLM-8B achieve mIoU scores of 16.92 and 6.82, respectively. On task-target grounding, both models outperform the strongest general-purpose baseline, achieving Joint@0.25 scores of 25.62 and 28.21, respectively, compared with 7.69.

These results indicate that stereo-representation training does not replace the monocular cues inherited from the pretrained model. Instead, StereoVLM remains usable when only one image is available, while the primary stereo evaluation measures the additional capability obtained from a valid left–right pair. We omit correspondence and disparity from this table because neither quantity is directly observable from a single image.

D.3 CALIBRATION-ROBUSTNESS DETAILS

We study robustness to residual errors in the relative extrinsics of the stereo cameras. Starting from a rectified pair, we keep the left image, query pixels, nominal focal length, nominal baseline, and left-view depth references fixed, and perturb only the pose of the right camera. The resulting right image is rendered under one of six single-axis perturbations: rotations R_x, R_y, and R_z, or translations t_x, t_y, and t_z. The experiment therefore measures sensitivity to imperfect stereo alignment without changing the scene or the queried left-view geometry.

The evaluation contains 400 base pairs, comprising 200 DROID pairs and 200 in-house robot pairs. Each pair contributes eight fixed left-image query points. For every axis, we construct three perturbation levels, giving 3,200 queried points per axis and level. The rotation magnitudes are shared across the three rotational axes, and the translation magnitudes are shared across the three translational axes. Averaged over the evaluation set, L1, L2, and L3 correspond to rotations of approximately $0.20°$, $0.59°$, and $0.98°$, or translations of 3.7 mm, 11.1 mm, and 18.4 mm. These magnitudes are selected per pair such that R_x and t_y induce mean vertical residuals of approximately 1, 3, and 5 pixels. The other axes use the same physical rotation or translation magnitude rather than being independently adjusted to produce the same image-space displacement.

Figure 7 illustrates the strongest perturbation for each axis. Each row shows the fixed left image, the perturbed right image, and the metric depth produced by FoundationStereo from that pair. The red point in the left image is the queried pixel. In the right image, the green circle marks its nominal correspondence and the red point marks its location after perturbation. The row labels report the mean horizontal and vertical correspondence displacement induced by the perturbation.

Rotations about x and translations along y primarily move correspondences away from the original epipolar row. By contrast, R_y and t_x primarily change horizontal position and therefore alter apparent disparity. The effects of R_z and t_z depend on image position and scene depth. FoundationStereo is run on the complete perturbed pair before the common display crop is applied. Its depth maps are included only as a qualitative diagnostic and do not contribute to the StereoVLM results in table 13. The black regions in the translation examples are invalid regions introduced by depth-based forward warping.

We evaluate three outputs. Correspondence EPE is measured against the right-image coordinate after perturbation. Direct depth is the metric depth predicted by StereoVLM in response to the depth question. Finally, we derive a depth estimate from the predicted right-image coordinate using

$$\hat{d}_i = x_i^l - \hat{x}_i^r, \qquad \hat{Z}_i^{\mathrm{corr}} = \frac{f_x B}{\hat{d}_i}, \tag{12}$$

where f_x and B are the nominal focal length and baseline. Because a perturbed pair does not in general satisfy the nominal rectified-stereo model, this quantity is a diagnostic of prediction under the original calibration rather than a physically corrected triangulation using the perturbed extrinsics. It is derived from the predicted correspondence and is not a separately predicted disparity output.

Table 13 gives the complete results for StereoVLM-2B. Correspondence error increases as the perturbation becomes stronger, from 2.48 pixels without perturbation to 4.93–7.13 pixels under the strongest perturbations. Direct metric depth is more stable. Under rotational perturbations, depth δ_1 remains between 90.3% and 91.1%, close to the nominal value of 90.5%. Translation has a larger but still moderate effect: at L3, depth δ_1 ranges from 86.1% to 87.8%, and MAE remains below 20 cm.

Depth derived under the nominal calibration follows the same broad trend, decreasing from 90.0% under the nominal condition to 88.0–89.2% under the strongest rotations and 84.6–85.5% under the strongest translations. Meanwhile, correspondence EPE increases consistently with perturbation magnitude. Thus, the model remains numerically stable but does not fully adapt its correspondence predictions to the perturbed geometry. Together with the single-view results in section D.2, this suggests that direct depth combines binocular evidence with learned monocular and nominal-stereo priors.

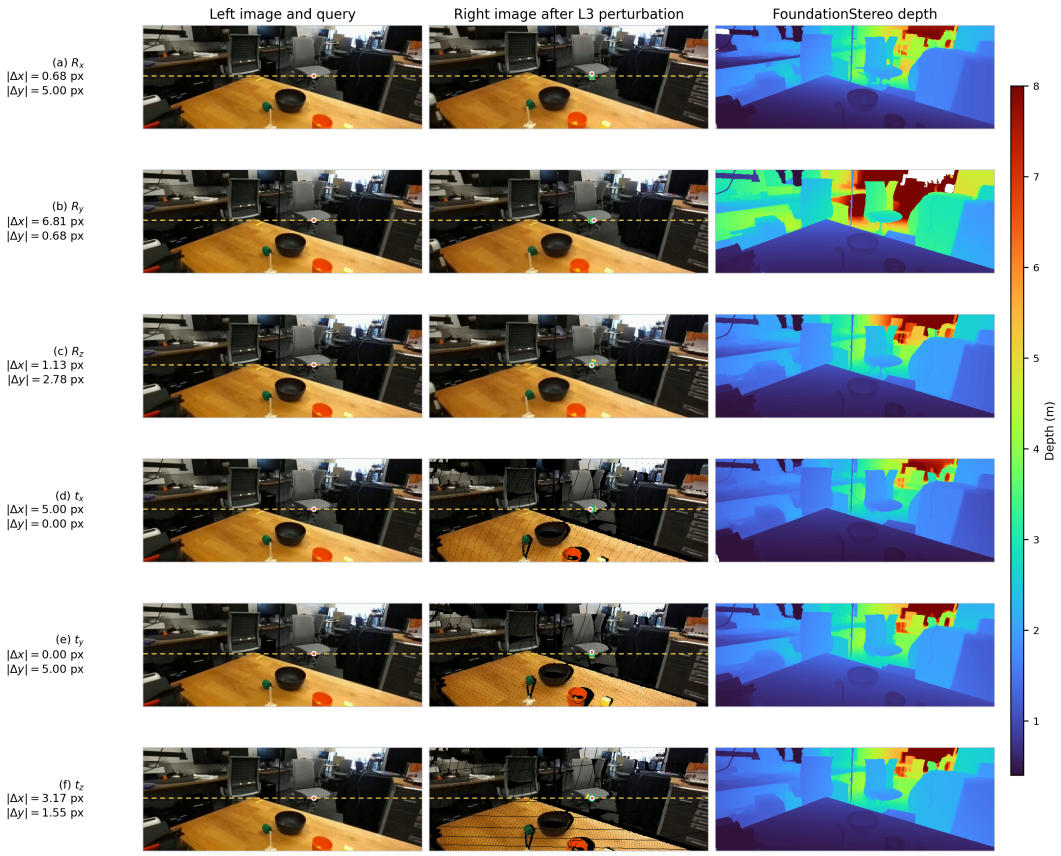

Figure 7: Qualitative effects of the strongest single-axis calibration perturbations. Each row shows the fixed left image and query, the right image after an L3 perturbation, and the metric depth produced by FoundationStereo from the perturbed pair. Green circles mark nominal correspondences, red points mark their locations after perturbation, and yellow arrows show the induced displacement. Row labels give the perturbation direction and its mean horizontal and vertical displacement. FoundationStereo receives the complete perturbed images; the common crop is applied only for display, and all depth maps share the same color scale. The displayed DROID example uses 0.99° rotations and 21.2 mm translations; perturbation magnitudes vary slightly by pair because they are calibrated from scene geometry.

D.4 Embodied-Control Transfer

We further examine whether stereo representation learning transfers to embodied control. For each VLM initialization, we train separate stereo and single-left policies under the same downstream training protocol. The stereo policy is evaluated with either the correct stereo pair or the left image duplicated into the right-view slot; these two results use the same policy checkpoint. The single-left policy is trained and evaluated using only the left image. table 14 reports average task success rates.

Without stereo representation training, the separately trained single-left policies outperform their stereo counterparts: 36.4% versus 34.9% for Qwen3-VL, and 39.5% versus 38.1% for StereoVLM-2B (+general). Thus, adding a right view does not by itself improve control under this training protocol.

With stereo representation training, the stereo policy reaches 42.4% success, the highest result in the table. This exceeds the separately trained single-left policy initialized from the same VLM by 2.4 percentage points. Because the policies are trained separately, this comparison measures the outcome of two training configurations, not the effect of removing the right view at inference time.

Table 13: Robustness of StereoVLM-2B to single-axis right-camera perturbations. Values within each perturbed row are reported in L1/L2/L3 order. Direct depth is predicted by StereoVLM. Nominal-calibration depth is derived from the predicted correspondence using $\hat{Z} = f_x B/(x_l - \hat{x}_r)$ without correcting for the perturbed extrinsics.

Condition	Correspondence	Direct Depth		Nominal Depth
	EPE px ↓	δ_1 ↑	MAE cm ↓	δ_1 ↑
Nominal	2.48	90.5	16.5	90.0
R_x, L1/L2/L3	2.91 / 4.54 / 6.41	90.6 / 90.7 / 90.6	16.2 / 16.5 / 17.3	90.2 / 89.8 / 89.2
R_y, L1/L2/L3	2.93 / 4.97 / 7.13	90.6 / 90.5 / 90.3	16.7 / 17.8 / 17.1	89.8 / 89.2 / 88.0
R_z, L1/L2/L3	2.69 / 3.65 / 4.93	90.6 / 91.1 / 91.0	16.4 / 16.4 / 16.2	89.2 / 89.4 / 89.1
t_x, L1/L2/L3	3.43 / 4.43 / 5.77	87.9 / 87.8 / 87.1	18.9 / 18.6 / 18.8	87.1 / 85.8 / 84.6
t_y, L1/L2/L3	3.67 / 5.14 / 7.00	88.6 / 88.1 / 87.8	18.7 / 18.1 / 19.3	86.2 / 86.7 / 85.5
t_z, L1/L2/L3	3.56 / 4.54 / 5.71	87.8 / 87.9 / 86.1	18.4 / 18.7 / 19.7	86.6 / 85.8 / 85.4

Table 14: Embodied-control success rates. Correct stereo and duplicated left are inference conditions for the same stereo-policy checkpoint; single left uses a separately trained policy. All policies follow the same downstream training and evaluation protocol.

VLM initialization	Correct stereo	Duplicated left	Single left
Qwen3-VL	34.9	31.6	36.4
StereoVLM-2B (general)	38.1	**36.8**	39.5
StereoVLM-2B (+stereo)	**42.4**	26.3	**40.0**

For the stereo-trained policy, replacing the right image with a copy of the left reduces success from 42.4% to 26.3%. This 16.1-point drop is a within-checkpoint intervention: it preserves the two-image input format while removing the valid stereo pairing. The result shows that this policy is sensitive to the content of the right-view slot, although the intervention alone cannot distinguish geometric dependence from other effects of an out-of-distribution input.

D.5 Qualitative Visual Results

fig. 8 compares dense depth predictions across DROID, Xperience-10M, and the held-out in-house benchmark Alongside the standard stereo-input result, we show the same StereoVLM-2B checkpoint with monocular input and the DepthVLM-4B baseline for reference.

fig. 9 visualizes representative semantic predictions under the standard stereo-input condition. Although only the rectified left image is shown for readability, both models receive the complete rectified left–right pair when producing each prediction. The three row groups correspond to object metric depth, semantic oriented 3D grounding, and task-target grounding.

For object depth, the yellow 2D box identifies the referred object and the label reports its predicted optical-axis depth in meters. For semantic 3D grounding, each oriented 3D box is projected into the rectified-left image. Task-target grounding additionally requires the model to identify the visible object specified by the manipulation instruction before predicting its 3D box. Green denotes the reference, orange denotes the StereoVLM-2B prediction, and blue denotes the StereoVLM-8B prediction.

In these examples, both model scales recover object depth closely and produce 3D boxes with similar centers, extents, and orientations to the references. The task-target examples further illustrate that the models can connect a manipulation instruction to the appropriate scene object and localize it in 3D. These examples are qualitative illustrations; aggregate performance and controlled input comparisons are reported in the main experiments.

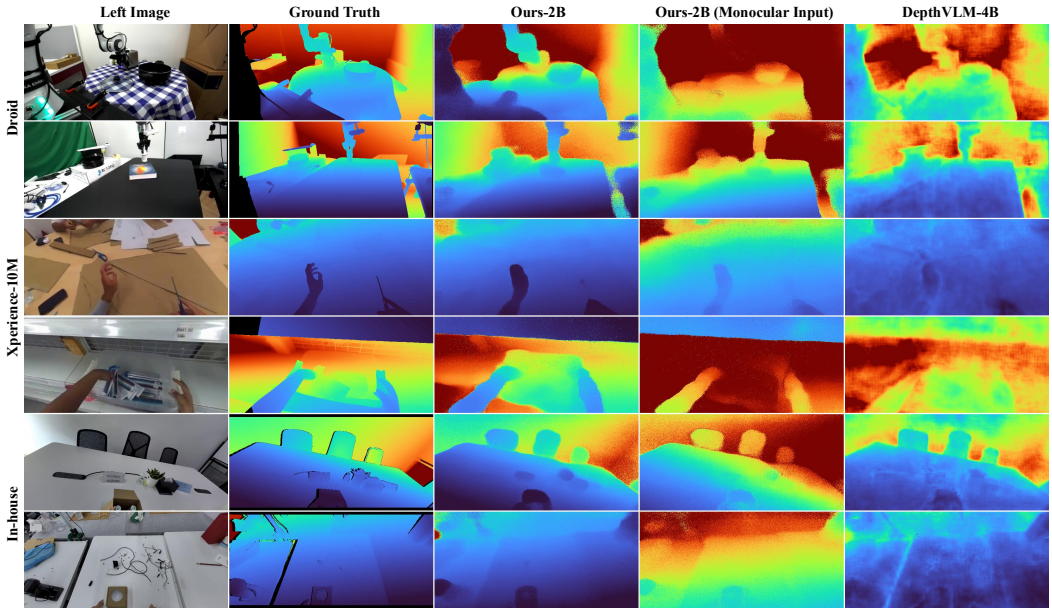

Figure 8: Qualitative dense depth results across three evaluation domains. Columns show the rectified left image, ground-truth depth, StereoVLM-2B with standard stereo input, the same StereoVLM-2B checkpoint with monocular input, and DepthVLM-4B. Rows are drawn from DROID, Xperience-10M, and the held-out in-house benchmark.

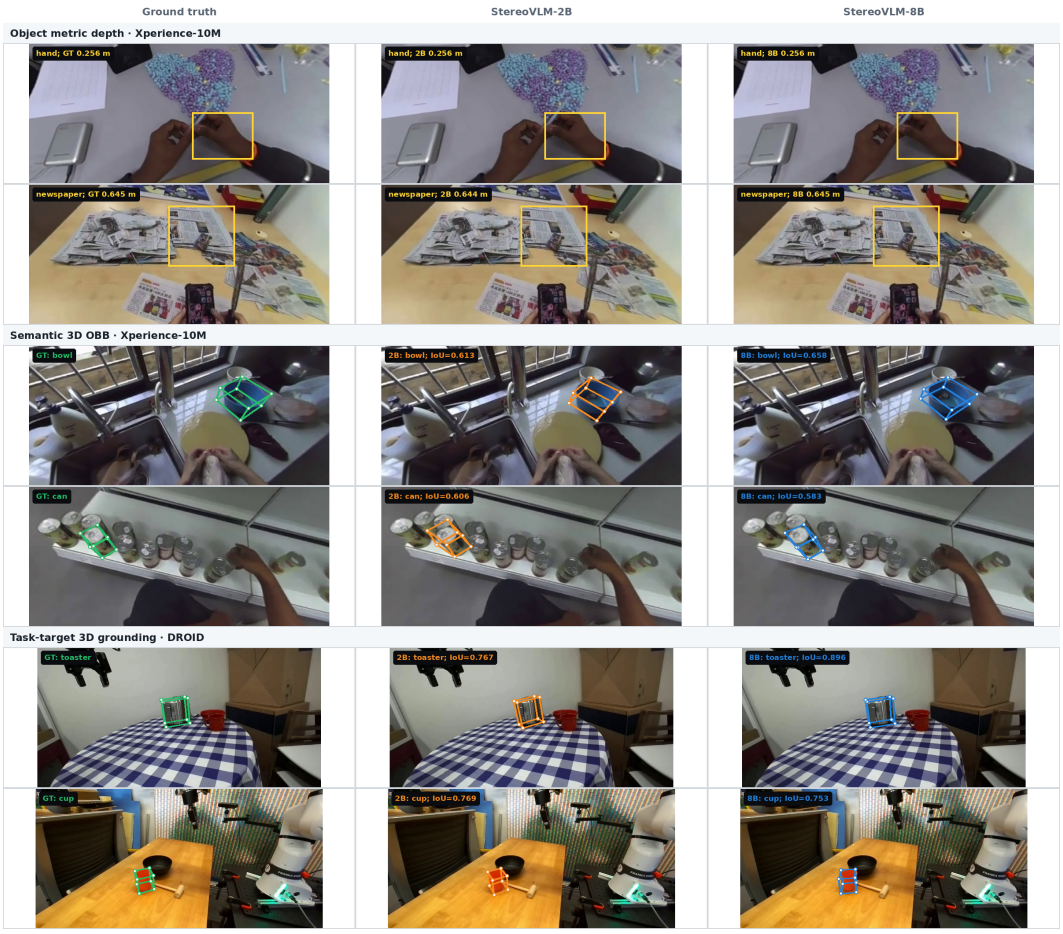

Figure 9: Qualitative semantic results under correct stereo input. Columns show the ground truth, StereoVLM-2B, and StereoVLM-8B. Object-depth examples report optical-axis depth in meters and use yellow boxes to indicate the queried objects. Semantic 3D OBB and task-target examples show oriented 3D boxes projected into the rectified-left image. Green, orange, and blue denote the ground truth, 2B prediction, and 8B prediction, respectively. Depth and semantic OBB examples are from Xperience-10M, while task-target examples are from DROID. Only the left-view visualization is displayed; predictions are produced from the complete stereo pair.