LESS THAN MEETS THE EYE: ACTION-SET IDENTIFIABILITY AND PRINCIPLED ABSTENTION FOR 3D PHYSICAL GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

A depth prediction can match its reference measurement and still describe a surface that no hand would touch. A corridor, its reflection in a planar mirror, and a photograph of that corridor on a screen can produce one image, while a hand stops in three places. Existing work addresses such scenes by enlarging what a predictor reports, through multi-layer formulations, detectors for mirrors and glass, and calibrated uncertainty. Each such quantity, however, describes a prediction or a predictor at a fixed viewpoint. None of them registers that the observation carries no evidence, because one image fits every explanation. We therefore ask a different question. Given a bounded set of camera motions, can an observer recover the physical explanation at all? Our approach, Intervene3D, calls two explanations separable to the extent that they predict different observations after the camera moves. A scene counts as identifiable once some available motion separates every remaining explanation, and the system abstains otherwise. A maximin rule then selects the motion separating the weakest contending pair. The experiments we conducted begin on a benchmark whose explanations are pixel-identical at the reference view, where abstention lowers false physical certainty from 0.832 to 0.000. Maximin selection there reaches a committed accuracy of 1.000 across four mechanisms, whereas a belief-weighted objective reaches 0.587 on mirrors. Turning to real photographs, our experiments on seven families reveal that all of them fail on 59.5 to 78.4% of images, while their confidence stays below chance of predicting these failures. Choosing the viewpoint that some hypothesis explains best then lowers ambiguous-region error by 31.3% on all thirteen checkpoints. Moreover, a gate over frozen encoder features reaches 0.887 out of fold and lowers the failure rate from 65.2% to 50.2% at 70% coverage. Asking what the available views can settle therefore precedes any answer about the physical geometry of the scene.

1 INTRODUCTION

A corridor viewed through an open doorway, reflected in a planar mirror, or displayed on a screen can produce an identical image. The three scenes differ in the surface a hand would touch, namely empty air, a pane of glass, and a rigid panel. An image therefore does not determine the physical explanation, and an answer from that image alone follows a prior rather than evidence. One change of viewpoint, by contrast, separates the three arrangements at once, because the surfaces send light along different paths as soon as the camera moves. Wherever perception drives action, physical contact sets the cost of such an error, and the quantity of interest becomes the first real surface along the ray. Figure 1 follows one real photograph through the decision proposed here.

Several lines of work already extend perception toward such configurations. Multi-layer formulations retain both a transparent interface and the scene behind that interface (Wen et al., 2025; Wen & Deng, 2026; Min et al., 2026), detectors localize mirrors and glass (Tan et al., 2021; Na et al., 2026), and selective prediction with split conformal prediction (Vovk et al., 2005; Lei et al., 2018) decides when a system should answer. Together these lines address the depth, the surfaces present, the next viewpoint, and the confidence of a prediction. The matched configurations above therefore raise a prior question. Can an observer recover the physical explanation at all? The question stands apart from accuracy and from uncertainty, which describe a prediction and a predictor. Identifiabil-

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.

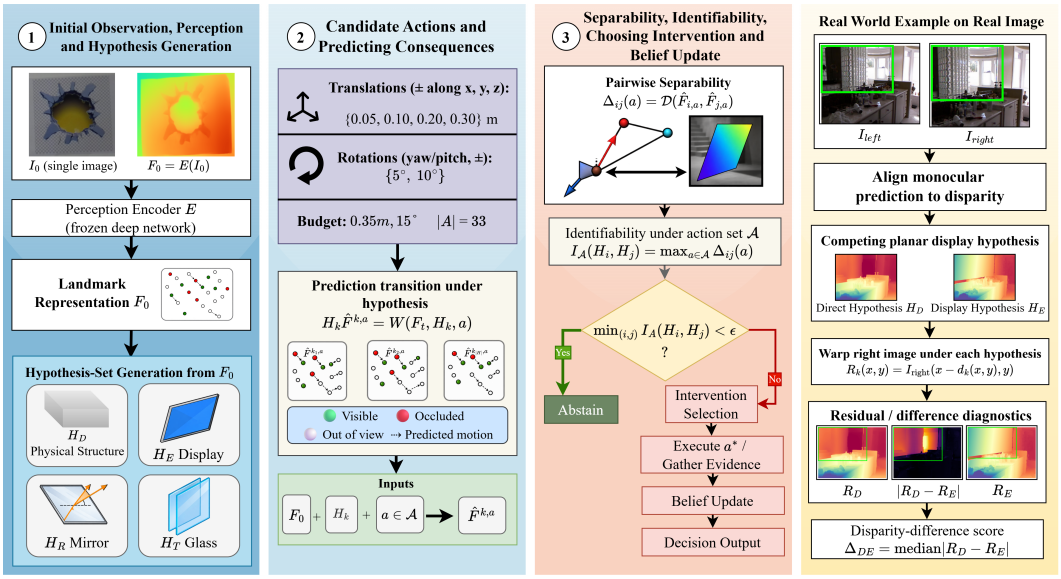


Figure 1: A high-level overview of Intervene3D on one real image. A stereo pair provides the executed action and the annotated region marks where the surface becomes ambiguous. The encoder geometry yields the direct hypothesis H_D and, after a plane fit over that region, the flat-panel hypothesis H_E. Warping the second view under each hypothesis produces the separability Δ_{DE} of equation 12, which equation 5 compares against the perceptual threshold ϵ to decide between answering and abstaining.

ity instead describes the triple formed by the competing explanations, the actions available to the observer, and the resolution at which differences become perceptible. For a view-tracked display observed within its aperture, the evidence contains no answer at all. Enlarging a model therefore changes nothing, whereas a suitable change of viewpoint resolves the case at once. We define identifiability by the actions an observer can execute.

Intervene3D evaluates identifiability before committing to an answer. Rather than a single geometry, the method carries four competing explanations of the same image, namely a scene physically present, content on a flat emissive panel, the virtual image behind a planar mirror, and a scene behind a transparent slab. All four agree at the reference view. For every motion in a bounded action set, a closed-form transition predicts what each explanation implies, and the distance between two predictions measures how far that motion separates the pair. Intervene3D answers only when the weakest contending pair clears a perceptual threshold, and the chosen motion maximizes that weakest separation. Our experiments support that framing across both settings. On the matched benchmark, a single-view classifier stays at exactly chance in every seed. Abstention then lowers false physical certainty from 0.832 to 0.000, and maximin selection reaches a committed accuracy of 1.000 across four mechanisms. On real photographs, deferring on identifiability lowers the failure rate among retained images from 65.2% to 53.1%, whereas deferring on the confidence of the predictor raises that rate to 76.6%. Choosing the best-explained viewpoint lowers ambiguous-region error by 31.3% on every checkpoint, and a gate over frozen encoder features lowers the failure rate to 50.2% at 70% coverage. The remainder of this work develops the four contributions set out below.

1. **Action-set identifiability and intervention selection.** We define separability under camera motion, and we identify a scene only when some available action separates every remaining hypothesis beyond a perceptual threshold. Actions then follow from maximin separability, known in optimal design as T-optimality.
2. **A benchmark and protocol with measured ambiguity.** Our cases remain pixel-identical at the reference view to order 10^{-14} px and include instances that no action resolves, which makes correct abstention measurable. Calibrated stereo provides the executed action.
3. **Evidence that the failure spans architectures and scenes.** Thirteen checkpoints from Depth Anything, DepthPro, MoGe, Metric3D, and UniDepth report depicted rather than physical geometry on 59.5 to 78.4% of illusion-benchmark images, and on 67.0% of 123 Mirror3D frames, while their confidence stays below chance at flagging these errors.
4. **A predictor-agnostic abstention gate.** A boosted policy predicts failures at 0.881 out of fold across all 13 held-out checkpoints.

2 METHOD

Intervene3D maintains a set of competing physical explanations for an observed image, predicts what each explanation implies for every candidate camera motion, and answers only when some available motion separates every explanation still in contention. Figure 7 presents the pipeline.

2.1 HYPOTHESES AND THE ACTION SET

Let I_0 denote an image at reference pose C_0. Let $F_0 = E(I_0)$ denote the geometry that a perception encoder E reports for that image, expressed as landmarks carrying image coordinates, depths, and visibility flags.

Definition 1 (Image-formation hypothesis family)**.** Rather than F_0 alone, Intervene3D keeps a four-member family $\mathcal{H} = \{H_D, H_E, H_R, H_T\}$. The observed structure is physically present under H_D, depicted on a flat emissive panel under H_E, the virtual image of a real scene in a planar mirror under H_R, or seen through a plane-parallel transparent slab under H_T. Each hypothesis places the first touchable surface elsewhere. All four remain *matched* at C_0, being pixel-identical there, and any separation gained after a motion therefore follows from the motion alone.

Against that family the observer acts. An action $a \in SE(3)$ applies a rigid camera motion as an explicit intervention $\mathrm{do}(C_1 = C_0 \cdot a)$. The do-operator replaces conditioning here, because the observer chooses the viewpoint. Actions come from a bounded set $\mathcal{A}$ of translations, rotations, and the null action, enumerated in Section 3.1. A finite action set allows evaluation of identifiability by enumeration before any motion, and abstention accordingly becomes possible before the camera moves.

2.2 PREDICTING THE CONSEQUENCE OF AN ACTION

Each hypothesis maps current geometry and a candidate action to a predicted observation,

$$\hat{F}^{\,k,a} \;=\; W(F_t,\, H_k,\, a)\,, \tag{1}$$

where F_t denotes the geometry at time t, H_k the k-th hypothesis, and $\hat{F}^{\,k,a}$ what H_k predicts after execution of a. The transition W applies one shared projection to every hypothesis and differs only in the world that each hypothesis asserts. Any difference between predictions therefore follows from the mechanism alone. The mechanisms remain planar because planarity keeps every transition closed form. Display content maps between views by the plane-induced homography $K'(R - t\mathbf{n}^\top/d)K^{-1}$. A planar mirror reflects exactly across its plane, and the camera pose transforms by the Householder map $\mathbf{R} \mapsto (I - 2\mathbf{n}\mathbf{n}^\top)\mathbf{R}$ with translation $-2d\mathbf{n}$. A planar slab, in turn, displaces content axially by $d\,(1 - 1/n)$. Closed form makes the counterfactuals matched, because reference views then coincide to numerical precision and divergence after an action cannot arise from approximation error. A curved surface admits no exact transition, and a learned transition would reintroduce the confound that the shared projection removes. The formalism itself, however, remains independent of that restriction (Appendix F.1).

2.3 SEPARABILITY, IDENTIFIABILITY, AND ABSTENTION

Pairwise separability measures the distance between the consequences that two hypotheses predict for the same action, where i and j index the pair and D denotes a composite distance,

$$\Delta_{ij}(a) \;=\; D\Big(\hat{F}^{\,i,a},\, \hat{F}^{\,j,a}\Big)\,, \tag{2}$$

$$D \;=\; \lambda_m D_{\text{motion}} + \lambda_o D_{\text{occlusion}} + \lambda_g D_{\text{geometry}} + \lambda_f D_{\text{feature}}. \tag{3}$$

The motion term measures the root-mean-square image-space displacement between the two predicted landmark sets over jointly visible landmarks, in pixels. The occlusion term counts landmarks whose predicted visibility differs, with visibility decided by z-buffering the geometry of each hypothesis. The geometry term measures the root-mean-square three-dimensional distance between back-projected landmarks, and the feature term measures a learned appearance distance that remains unused here. The quantity of interest then becomes the best separability that the action set

attains. A pair counts as separable when any permitted motion distinguishes that pair, and the system abstains in the absence of such a motion,

$$\mathcal{I}_{\mathcal{A}}(H_i, H_j) \;=\; \max_{a \in \mathcal{A}} \; \Delta_{ij}(a), \tag{4}$$

$$\min_{(i,j)} \; \mathcal{I}_{\mathcal{A}}(H_i, H_j) \;<\; \epsilon, \tag{5}$$

with the minimum running over pairs still in contention and ϵ a perceptual threshold. The order of the two operators matters. A maximum over actions inside a minimum over pairs licenses an answer only when some action separates every undecided pair, and the pair left unseparated marks exactly where an error would occur.

2.4 Choosing the Intervention

We compare two objectives. The first objective sums the pairwise separabilities, in the manner that active perception commonly adopts,

$$a^{\star}_{\text{sum}} \;=\; \arg\max_{a \in \mathcal{A}} \; \sum_{i<j} p_i \, p_j \, \Delta_{ij}(a), \tag{6}$$

where the posteriors p_i, p_j let still-plausible pairs contribute more. Such a sum, however, can reach its maximum at an action that separates pairs outside contention while leaving the deciding pair unseparated. The second objective therefore replaces the sum with the weakest contending pair,

$$a^{\star}_{\text{maximin}} \;=\; \arg\max_{a \in \mathcal{A}} \; \min_{(i,j) \in \mathcal{C}} \; w_{ij} \, \Delta_{ij}(a), \tag{7}$$

where $\mathcal{C}$ collects pairs whose joint posterior mass has not collapsed and w_{ij} weights each pair by that mass. The form follows from the abstention rule, because a decision requires every contending pair separated, which makes the smallest separability the binding quantity. Optimal design for model discrimination knows this criterion as T-optimality and the present work transfers the criterion to active vision. The two objectives differ only where at least two hypothesis pairs exist, as in the six pairs of the synthetic benchmark and not in a stereo pair. Selection on real photographs accordingly runs through the belief update of equation 9. Among viewpoints that the observer could occupy, the rule prefers the one at which some hypothesis best accounts for the view obtained,

$$a^{\star}_{\text{evidence}} \;=\; \arg\min_{a \in \mathcal{A}} \; \min_{k} \; e_k(a), \tag{8}$$

with $e_k(a)$ the residual of hypothesis k after execution of a. Both rules act on the same machinery, one through the separability entering the belief and the other through the residual leaving that belief.

2.5 Belief Update and Output

After execution of the selected action and observation of O', each hypothesis carries the residual $e_k \;=\; D(\hat{F}^{\,k,a^{\star}}, E(O'))$ under the same distance equation 3. Prediction error therefore shares a scale with separability. The belief updates multiplicatively,

$$p_{t+1}(H_k) \;\propto\; \exp(-\beta \, e_k) \; p_t(H_k), \tag{9}$$

where β converts pixel error into log-likelihood. A posterior floor meanwhile stops one unlucky observation from eliminating a hypothesis permanently. The system then names the mechanism and reports the contact geometry, the first physical surface along each ray, the glass for a mirror and the panel for a display. The system commits only when the posterior maximum exceeds a threshold τ.

2.6 Protocol for Real Photographs

Real data requires no change to Intervene3D, because a calibrated rectified stereo pair already constitutes an executed action with known parameters. Real photographs therefore have $|\mathcal{A}| = 1$. The setting demands only that no step absorb the effect under measurement. A monocular predictor returns relative inverse depth, defined up to an unknown scale and shift. We fit both quantities by least squares against reference disparity outside the ambiguous region, and we then apply them inside that region, which yields the disparity predicted under H_D,

$$d_D \;=\; s\,\hat{p} + t, \tag{10}$$

where $\hat{p}$ denotes the raw prediction, s the scale, and t the shift. Fitting equation 10 where the predictor performs well, and scoring where the predictor performs poorly, stops those two parameters from absorbing the discrepancy under study. The competing flat-panel hypothesis is the least-squares plane through that aligned prediction, fitted to the prediction and not to the reference, because a hypothesis built from the answer would place ground truth inside a test-time quantity. We then evaluate separability in the image. Rectification aligns the pair, hence column x with disparity d corresponds to column $x - d$ in the other view. The right photograph warps into the left frame under each hypothesis, and we compare the two warps,

$$R_k(x, y) \;=\; I_{\text{right}}(x - d_k(x, y),\; y)\,, \qquad k \in \{D, E\}, \tag{11}$$

$$\Delta_{DE} \;=\; \text{median}\,|R_D - R_E|\,, \tag{12}$$

with d_k the disparity that hypothesis k predicts and the median taken over pixels where both warps remain valid. Warping the same photograph twice cancels illumination, sensor noise, and rectification error. The result carries intensity units rather than the pixels of equation 3, and a disagreement counts only where the image carries structure to reveal that disagreement. A featureless region, where the hypotheses predict different geometry and identical pixels, then enters the record as unidentifiable at any baseline, whereas a disparity-space measure would call the same region highly identifiable. Two further choices complete the protocol. A physical sheet or panel is flat, and we therefore retain an image only when a plane fitted to the reference disparity inside the region attains $R^2 \geq 0.90$. Below that value the stereo reference has locked onto displayed content rather than onto the panel. Failure follows a definition independent of every signal under test. A predictor fails when its relative error inside the region exceeds its error outside that region, and such a label mentions neither confidence nor identifiability. Every value of the area under the receiver operating characteristic curve (AUROC) follows the convention that a larger value indicates a more likely error.

3 Results

3.1 Experimental Setup

The synthetic benchmark instantiates each base scene under every mechanism and adjusts content until the reference views coincide to order 10^{-14} px, which makes chance-level single-view performance a verified property of the data. The benchmark also retains configurations that no permitted action resolves, which makes correct abstention measurable. Splits follow base scenes throughout. We report causal explanation accuracy (CEA), committed accuracy and the false physical certainty rate (FPCR), all defined in Table 4.

The action set contains translations of $\{0.05, 0.10, 0.20, 0.30\}$ m along three axes, rotations of $\{5^\circ, 10^\circ\}$ in yaw and pitch, and the null action, which yields $|\mathcal{A}| = 33$. The composite distance of equation 3 runs at $\lambda_m = 1$, $\lambda_o = 4.0$ px, and $\lambda_g = \lambda_f = 0$. Priors stay uniform at $t = 0$, and the three remaining constants stay at $\epsilon = 1.0$ px, $\beta = 1.0$, and $\tau = 0.45$ without tuning. Appendices E.2, E.1 and E.3 sweep each one.

We experimented with thirteen models spanning seven architecture families and a $38\times$ range in parameter count, without training or tuning on any evaluation set. Every model stays frozen, and we fit only a policy on top, under folds split by base scene, hyperparameters chosen on inner folds, and a leave-one-model-out protocol (Appendix D.1).

3.2 Synthetic Benchmark

Table 1 reports Phase 1 over 11 seeds with the benchmark redrawn per seed, and the intervals therefore reflect the data-generating process. The single-view classifier stays at exactly chance in every seed, which confirms that the matched construction leaves nothing for one view to exploit. Moreover, the ordering from single-view to passive to intervention-aware holds strictly in 11 of 11 seeds. A second study then isolates the choice of intervention. The summed objective equation 6 reaches 0.587 and 0.956 committed accuracy on mirrors, whereas maximin selection equation 7 reaches 1.000 in each case (Figure 2c). Across four mechanisms, maximin selection also exceeds the passive baseline by $+0.297$ ($t = 41.6$, 11 of 11 seeds). A belief-weighted sum can reach its maximum at an action that separates pairs outside contention while leaving the deciding pair

Table 1: Phase 1 of the synthetic benchmark, over three mechanisms, eleven seeds, chance at 0.333. CEA counts abstention as an error, and committed accuracy scores the answered scenes alone.

Method	CEA	Committed	Abstention	AUROC	FPCR
Single-view classifier	0.333±0.000	0.333±0.000	0%	0.505±0.066	0.000±0.000
Passive multi-view	0.621±0.046	0.621±0.046	0%	0.748±0.065	0.868±0.107
Generic next-best-view proxy	0.371±0.050	0.706±0.029	48%	0.898±0.026	0.000±0.000
Random intervention	0.409±0.056	0.730±0.027	44%	0.925±0.024	0.000±0.000
Max-baseline intervention	0.502±0.061	0.824±0.019	39%	0.963±0.028	0.000±0.000
ablations					
no hypothesis conditioning	0.000±0.000	—	100%	0.500±0.000	0.000±0.000
noisy encoder	0.557±0.034	0.848±0.049	34%	0.994±0.016	0.008±0.026
no abstention	**0.722±0.029**	0.722±0.029	0%	1.000±0.000	0.832±0.107
Intervene3D (summed)	0.560±0.032	0.855±0.054	34%	1.000±0.000	0.000±0.000
Intervene3D (maximin)	0.563±0.076	**1.000±0.000**	44%	0.927±0.032	**0.000±0.000**

Table 2: Thirteen checkpoints on 123 annotated frames of real mirrors (Tan et al., 2021). The last two columns carry relative error inside against outside the mirror region and the share of frames placing the first surface behind the glass.

Checkpoint	Family	Fooled	Error inside / outside	Surface behind glass
DA-V2-Small	Depth-Anything-V2	65.0%	1.72×	32.5%
DA-V2-Base	Depth-Anything-V2	71.5%	2.00×	30.1%
DA-V2-Large	Depth-Anything-V2	65.0%	1.82×	27.6%
DA-transparency	Depth-Anything-V2	41.5%	0.76×	30.1%
DA3-Small	Depth-Anything-3	**79.7%**	2.85×	59.3%
DA3-Base	Depth-Anything-3	76.4%	3.38×	56.9%
DA3-Large	Depth-Anything-3	75.6%	2.57×	42.3%
DepthPro	DepthPro	65.0%	1.91×	28.5%
MoGe-2	MoGe-2	74.8%	2.19×	22.8%
MoGe-3-ViT-L	MoGe-3	73.2%	2.09×	23.6%
Metric3D-Small	Metric3D	78.9%	**5.47×**	**66.7%**
Metric3D-Large	Metric3D	41.5%	0.82×	32.5%
UniDepth-v2	UniDepth	62.6%	1.93×	19.5%

unseparated. An action scoring zero on a pair cannot separate that pair at any ϵ, and no choice of threshold closes the gap.

3.3 Real Photographs

We evaluate thirteen published checkpoints spanning seven architecture families and a 38× range in parameter count, all as released (Yang et al., 2024; Lin et al., 2026; Bochkovskiy et al., 2025; Wang et al., 2025; Kong et al., 2026; Yin et al., 2023; Piccinelli et al., 2024; 2026) (Figure 2a,b). Every checkpoint fails on 59.5 to 78.4% of admissible images, with the newest family among the most affected. The confidence proxy taken from each predictor stays below chance at predicting its own failure, with values from 0.234 to 0.465. Identifiability, by contrast, ranges from 0.467 to 0.719 and exceeds confidence on all thirteen. Calibration cannot repair a signal ordered against the quantity of interest. Both signals meet split-conformal coverage at all four levels. Deferring on confidence, however, leaves a retained set failing at 76.6% against a base rate of 65.2%, whereas identifiability lowers the same quantity to 53.1% (Figure 2d). The criterion of Section 2.6 excludes 159 of 455 images whose stereo reference locks onto displayed content. Beyond this benchmark, two independent labels agree. Accuracy split by the per-point layer index of a second benchmark reaches 0.848 against 0.737. Table 2 then repeats the evaluation on 123 photographs of real mirrors whose contact geometry carries an independent annotation (Tan et al., 2021). The same thirteen checkpoints report the depicted scene at 67.0% there, against 66.8% on displays. Relative error inside the mirror region reaches 5.47× the error outside, and up to two thirds of frames place the first surface behind the glass. A different optical mechanism, annotated by a different group against a human-placed plane rather than a stereo reference, therefore produces the same rate.

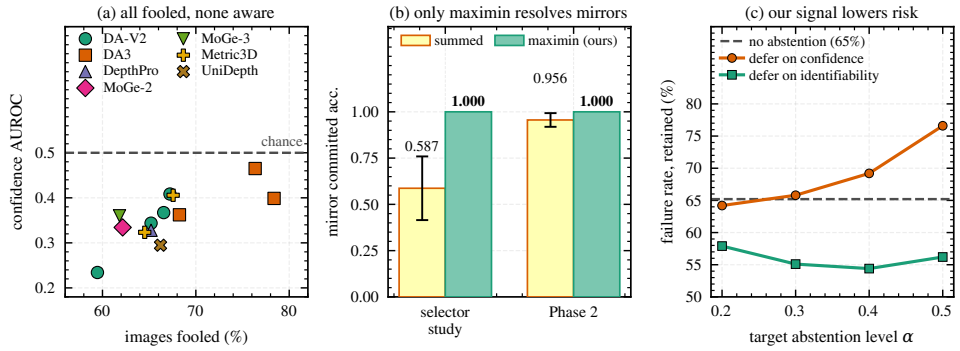


Figure 2: Four views of the evaluation over the thirteen checkpoints. (a) Images fooled. (b) Confidence AUROC. (c) Committed accuracy on mirrors. (d) Split-conformal selective risk.

Table 3: The three categories of the illusion benchmark (Yao et al., 2025), over thirteen checkpoints. Median Δ is the median separability, and the last column counts images falling below the perceptual threshold ϵ.

Category	**Images**	**Fooled**	**Range**	**At** 100%	**Error in/out**	**Median** Δ	$\Delta < \epsilon$
Printed sheet on a table	97	43.0%	23.7–76.3	0/13	0.91×	1.07	82.5%
Poster beside an object	104	58.9%	34.6–75.0	0/13	1.23×	0.57	86.5%
Monitor in the frame	95	**99.9%**	98.9–100	**12/13**	2.62×	**0.25**	**100%**
All admissible	**296**	**66.8%**	59.5–78.4	—	1.66×	0.45	89.5%

3.4 Qualitative and Error Analysis

Figure 3 places all thirteen checkpoints on one photograph of a poster beside a window frame. The stereo reference reports a near planar surface across the outlined region, whereas every checkpoint reports the structure painted on the poster.

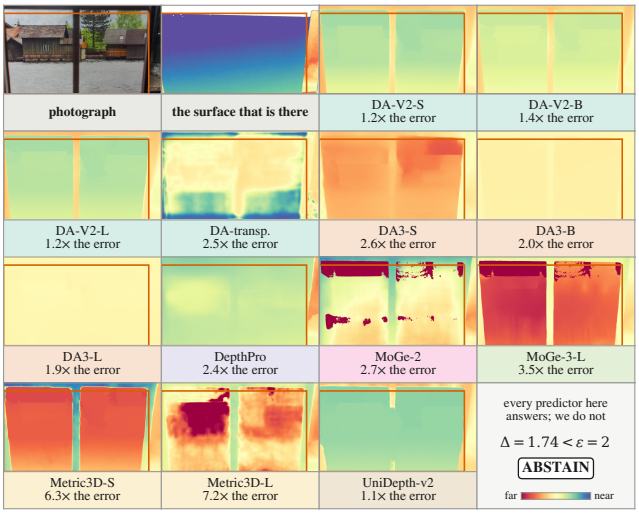


Figure 3: All thirteen checkpoints on one photograph of a poster beside a window frame, with the ambiguous region outlined and the stereo reference beside each prediction.

There the failure stays local, at 1.1 to 7.2 times the error each checkpoint makes outside. The aggregate rate above averages over three conditions that behave differently, and Table 3 separates them. On the 97 images of a sheet lying on a table, failure remains a minority outcome at 43.0%. The median separability of 1.07 leads the three categories, and the relative error inside the ambiguous region stays below the error outside that region at 0.91×. On the 95 images containing a monitor, failure instead becomes total, and twelve of the thirteen checkpoints fail on *every* image. The relative error inside the region reaches 2.62×, the median separability falls to 0.25, and every image lies below ϵ. Failure therefore concentrates exactly where the executed action separates the two hypotheses least, which is the pattern predicted by the criterion of Section 2.3.

Figure 4 presents three images with their predictions, namely a monitor in the frame, a sheet lying on a table, and a poster beside an object. In each case the reference geometry contains a flat panel across the annotated region, and the prediction follows the scene painted on that panel. The top row carries the most confident prediction among all 296 images, which places the largest confidence of

the evaluated set on a surface that a hand would meet at the panel. Separability falls below ϵ in all three cases, which makes equation 5 decline to answer where the predictor commits.

The residual errors of Intervene3D follow the same geometry. Where a sheet lies at an oblique angle and the executed baseline moves across that obliquity, separability rises above ϵ, the system commits, and the commitment agrees with the reference. Where the panel fills the frame and the stereo baseline shifts along the plane of the panel, no hypothesis leaves a photometric trace, and the abstention that follows records the absence of evidence rather than a mechanism.

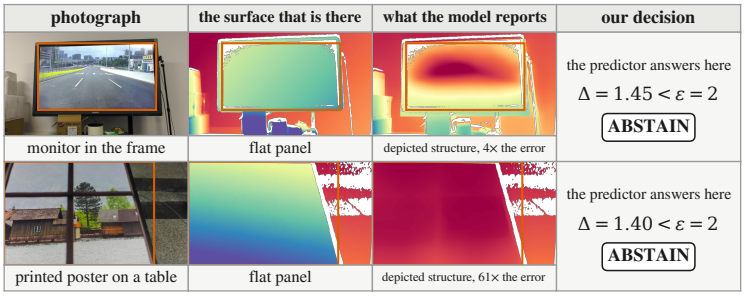


Figure 4: Three photographs with the reference surface beside each prediction, namely a monitor in the frame, a sheet lying on a table, and a poster beside an object.

What Separability Measures. Both hypotheses come from the output of the predictor itself, and a direct control therefore tests whether the quantity describes the scene or that one network (Figure 5). Scoring the separability of one checkpoint against the failures of a different checkpoint, over all 156 ordered pairs, reaches **0.637**. The cross-model score stays above chance on every pair and above the 0.597 that each checkpoint attains on its own failures. Idiosyncrasy does not transfer across networks, whereas scene structure does. A separability built from one predictor consequently describes the scene rather than the predictor itself.

3.5 Choosing Where to Look on Real Photographs

The benchmark photographs each static scene from several camera positions. The median stereo disparity differs across frames while the scene does not, which is what a moved observer produces. Fifty-five scenes therefore offer a real action set of three or more viewpoints. We apply equation 8 and measure error inside the ambiguous region, which makes the outcome independent of every other part of the image. The best-explained viewpoint then lowers that error by 31.3% against a random choice, on 13 of 13 checkpoints, and the fooled rate falls from 48.0% to 44.1%, on 9 of 13. Two further baselines place that gain in context. Holding the first frame of each scene, which amounts to no choice at all, leaves the error at 0.0671. A uniformly random choice reaches 0.0591, a viewpoint chosen by the confidence of the predictor reaches 0.0439, and the best-explained observation reaches 0.0406. Selecting on the evidence thus improves on selecting by confidence, and the ordering across the four rules follows the amount of evidence each rule consults. Real images carry no mechanism annotation, and this experiment measures error inside the ambiguous region at the viewpoint that the rule selects.

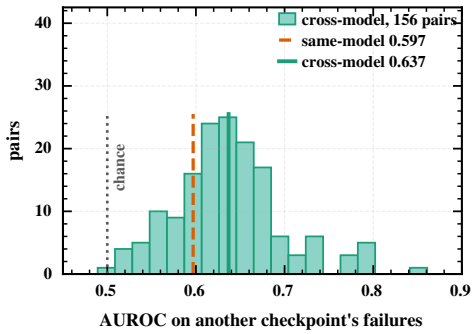


Figure 5: Separability from one checkpoint scored against the failures of a *different* checkpoint, over all 156 ordered pairs.

3.6 Abstention from the Frozen Features of the Predictor

The signals above are hand-designed, which raises the question of whether a learned combination performs better and whether the learned quantity remains specific to one predictor. Under the leave-one-model-out protocol, a gradient-boosted policy trained on twelve checkpoints and tested on the thirteenth reaches 0.881 against 0.768 for the strongest single feature, ahead on 13 of 13. Moreover, on a single checkpoint a head over frozen encoder features reaches 0.887 out of fold, with the layer

chosen on inner folds and consistently an early one. No parameter of the encoder changes, and the gate consequently adds no retraining and no architectural change. In practice, answering the most confident 70% of inputs lowers the failure rate from 65.2% to 50.2%, and answering half lowers the same rate to 37.8%. Every threshold comes from out-of-fold scores under folds split by scene. The curve stays monotone across the whole range, from 61.3% at 90% coverage to 23.6% at 30%. A deployment thus trades answered inputs against failures at a rate the operator chooses rather than at one fixed point.

4 Related Work

Prior work localizes mirrors and glass as scene elements in their own right (Yang et al., 2019; Mei et al., 2020; Tan et al., 2021), models transparency through decomposed radiance (Na et al., 2026), and quantifies the difficulty with dedicated benchmarks (Zama Ramirez et al., 2024; Chen et al., 2022; Dai et al., 2022). Robustness studies carry the same concern to corrupted inputs (Kong et al., 2023), and a recent challenge gathers current systems on high-resolution depth for exactly these materials (Zama Ramirez et al., 2025). Multi-layer formulations then represent more than one surface along a ray (Wen et al., 2025; Wen & Deng, 2026; Min et al., 2026), and an illusion benchmark documents at scale that predictors report depicted rather than physical structure (Yao et al., 2025). This work steps back from recovering that geometry and asks first whether the available observations determine the mechanism behind an appearance.

Classical analyses establish that image evidence alone can leave a family of scenes indistinguishable (Kutulakos & Steger, 2008), while active and next-best-view perception selects viewpoints by a belief-weighted sum over competing possibilities (Yang & Waslander, 2022; Rainforth et al., 2024). We ask instead whether any view within a bounded action set licenses a commitment. The sequential design of experiments already treats the choice of a next experiment as a decision between rival hypotheses. From that lineage we borrow the maximin criterion that optimal design calls T-optimality, together with the interventional framing of causal identifiability (Lippe et al., 2022). Deferral rules (Geifman & El-Yaniv, 2017; 2019) and split conformal prediction (Vovk et al., 2005; Lei et al., 2018) constrain how many predictions a system retains rather than which ones. A calibrated procedure therefore inherits the ranking of the underlying score rather than repairing that ranking.

5 Conclusion

In this paper, we investigate whether a physical explanation is recoverable before asking what that explanation is, and the measurements support that order. On the synthetic benchmark, whose competing mechanisms are matched at the reference view, a single-view classifier stays at exactly chance in every seed. Abstention then lowers false physical certainty from 0.832 to 0.000, and selecting the action that separates the weakest contending pair reaches a committed accuracy of 1.000 on all four mechanisms. On real photographs, thirteen checkpoints across seven architecture families report the depicted scene on 59.5 to 78.4% of images, and their confidence stays below chance at predicting these errors. Split conformal prediction meanwhile meets its coverage guarantee while retaining a set that fails more often than answering everywhere. Three further measurements indicate where the missing evidence can come from. The layer labels of an independent benchmark yield 0.848 against 0.737, and the best-explained viewpoint lowers ambiguous-region error by 31.3% on every checkpoint. A boosted policy then reaches 0.881 out of fold across all thirteen unseen architectures, and a frozen-encoder gate lowers the failure rate from 65.2% to 50.2% at 70% coverage while the predictor stays unchanged.

Limitations. Intervene3D requires an executed action, the hypothesis family covers planar optics alone, and real photographs provide a single stereo baseline. Appendix F.2 states the scope in full, and Appendix F.1 describes the route beyond planar optics.

References

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second.

In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 75602–75637, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bc8b2058fd96978a4146f18298cb2d39-Abstract-Conference.html.

Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. ClearPose: Large-scale transparent object dataset and benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 381–396, 2022. URL https://doi.org/10.1007/978-3-031-20074-8_22.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 374–391, 2022. URL https://doi.org/10.1007/978-3-031-19842-7_22.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. RoboDepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 21298–21342, 2023. URL https://doi.org/10.52202/075280-0932.

Lingyu Kong, Ruicheng Li, Ruicheng Wang, Sicheng Xu, Chengtang Yao, Jianfeng Xiang, and Jiaolong Yang. MoGe-3: Fine-detail monocular geometry estimation with self-guided sparse volumetric refinement. *arXiv preprint arXiv:2607.17967*, 2026. URL https://arxiv.org/abs/2607.17967.

Kiriakos N. Kutulakos and Eron Steger. A theory of refractive and specular 3D shape by light-path triangulation. *International Journal of Computer Vision*, 76(1):13–29, 2008. URL https://doi.org/10.1007/s11263-007-0049-9.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. URL https://doi.org/10.1080/01621459.2017.1307116.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=yirunib8l8.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162, pp. 13557–13603, 2022. URL https://proceedings.mlr.press/v162/lippe22a.html.

Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W. H. Lau. Don't hit me! glass detection in real-world scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3684–3693, 2020. URL https://doi.org/10.1109/CVPR42600.2020.00374.

Junhong Min, Jimin Kim, Minwook Kim, Cheol-Hui Min, Youngpil Jeon, and Minyong Choi. DepthFocus: Controllable depth estimation for see-through scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12595–12605, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Min_DepthFocus_Controllable_Depth_Estimation_for_See-Through_Scenes_CVPR_2026_paper.html.

Youngju Na, Jaeseong Yun, Soohyun Ryu, Hyunsu Kim, Sung-Eui Yoon, and Suyong Yeon. GLINT: Modeling scene-scale transparency via gaussian radiance transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7256–7265, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Na_GLINT_Modeling_Scene-Scale_Transparency_via_Gaussian_Radiance_Transport_CVPR_2026_paper.html.

Hoang Nguyen, Xiaohao Xu, and Xiaonan Huang. The 3D mirage: Probing and taming 3D hallucinations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2512.15423.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116, 2024. URL https://doi.org/10.1109/CVPR52733.2024.00963.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2354–2367, 2026. URL https://doi.org/10.1109/TPAMI.2025.3628473.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. URL https://doi.org/10.1214/23-STS915.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 746–760, 2012. URL https://doi.org/10.1007/978-3-642-33715-4_54.

Xiangyu Sun, Shijie Wang, Fengyi Zhang, Lin Liu, Caiyan Jia, Ziying Song, Zi Huang, and Yadan Luo. VGGT-World: Transforming VGGT into an autoregressive geometry world model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 382–400, 2026. URL https://doi.org/10.1007/978-3-032-37271-0_21.

Jiaqi Tan, Weijie Lin, Angel X. Chang, and Manolis Savva. Mirror3D: Depth refinement for mirror surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15985–15994, 2021. URL https://doi.org/10.1109/CVPR46437.2021.01573.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. URL https://doi.org/10.1007/b106715.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5261–5271, 2025. URL https://doi.org/10.1109/CVPR52734.2025.00496.

Hongyu Wen and Jia Deng. SeeGroup: Multi-layer depth estimation of transparent surfaces via self-determined grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7299–7309, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wen_SeeGroup_Multi-Layer_Depth_Estimation_of_Transparent_Surfaces_via_Self-Determined_Grouping_CVPR_2026_paper.html.

Hongyu Wen, Yiming Zuo, Venkat Subramanian, Patrick Chen, and Jia Deng. Seeing and seeing through the glass: Real and synthetic data for multi-layer depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6715–6725, 2025. URL https://doi.org/10.1109/ICCV51701.2025.00632.

Jun Yang and Steven L. Waslander. Next-best-view prediction for active stereo cameras and highly reflective objects. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3684–3690, 2022. URL https://doi.org/10.1109/ICRA46639.2022.9811917.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 21875–21911, 2024. URL https://doi.org/10.52202/079017-0688.

Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W. H. Lau. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8808–8817, 2019. URL https://doi.org/10.1109/ICCV.2019.00890.

Chengtang Yao, Zhidan Liu, Jiaxi Zeng, Lidong Yu, Yuwei Wu, and Yunde Jia. 3D visual illusion depth estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 86985–87017, 2025. URL https://doi.org/10.52202/085713-2622.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9009–9019, 2023. URL https://doi.org/10.1109/ICCV51070.2023.00830.

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*, 46(1): 85–102, 2024. URL https://doi.org/10.1109/TPAMI.2023.3323858.

Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, Samuele Salti, Stefano Mattoccia, Zhe Zhang, Yang Yang, Wu Chen, Anlong Ming, Mingshuai Zhao, Mengying Yu, Shida Gao, Xiangfeng Wang, Feng Xue, Jun Shi, Yong Yang, Yong A, Yixiang Jin, Dingzhe Li, Aryan Shukla, Liam Frija-Altarac, Matthew Toews, Hui Geng, Tianjiao Wan, Zijian Gao, Qisheng Xu, Kele Xu, Zijian Zang, Jameer Babu Pinjari, Kuldeep Purohit, Mykola Lavreniuk, Jing Cao, Shenyi Li, Kui Jiang, Junjun Jiang, and Yong Huang. NTIRE 2025 challenge on HR depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 978–992, 2025. URL https://doi.org/10.1109/CVPRW67362.2025.00098.

SUPPLEMENTARY MATERIAL

CONTENTS

A Experimental Details **13**
A.1 Why One Image Admits Several Mechanisms . . . 13
A.2 The Decision Pipeline . . . 14
A.3 The Real-Photograph Protocol in Detail . . . 14
A.4 Minimum Causal Resolving Baseline . . . 15
A.5 Metric Definitions . . . 16

B The Synthetic Benchmark **16**
B.1 Choosing the Intervention, in Full . . . 16
B.2 Calibrated Deferral and Imperfect Execution . . . 16

C Real Photographs **17**
C.1 Per-Checkpoint Results . . . 17
C.2 Extended Results by Architecture Family . . . 17
C.3 Choosing Where to Look . . . 18
C.4 Second Dataset with Ordinal Supervision . . . 18
C.5 Real Mirrors, with Annotated Contact Geometry . . . 19

D Learned Abstention **20**
D.1 Learned Abstention . . . 20
D.2 Full Leave-One-Model-Out Results . . . 21
D.3 Full Abstention Retention Curve . . . 22

E Robustness and Controls **22**
E.1 Threshold Sensitivity . . . 22
E.2 Sensitivity to the Occlusion Weight . . . 23
E.3 Threshold, Temperature and Sensor Scale . . . 23
E.4 Comprehensive Analyses . . . 25

F Scope, Limitations and Reproducibility **28**
F.1 Extending Beyond Planar Optics . . . 28
F.2 Scope and Limitations . . . 28
F.3 Reproducibility . . . 29
F.4 Use of Generative Artificial Intelligence . . . 29

A EXPERIMENTAL DETAILS

In the following sub-sections we further describe the method and experimental details.

A.1 WHY ONE IMAGE ADMITS SEVERAL MECHANISMS

Figure 6 presents the premise of the paper in one picture, namely three physical mechanisms that produce the same observed image, together with the effect that an executed motion has on each mechanism. Under the direct hypothesis H_D the camera views real three-dimensional structure spread over several depths, under the display hypothesis H_E the same content lies on a single plane, and under the mirror hypothesis H_R the camera views a virtual scene behind a reflective surface. The light arriving at the sensor stays the same in all three cases, so the reference image alone cannot favor any of them (panel 1). Motion breaks the tie because each mechanism moves its content by a different law (panel 2). Under a downward translation of the camera, content at several depths shifts in the image by an amount that depends on its distance, content on a display moves rigidly with the plane that carries it and gains no parallax of its own, and content behind a mirror follows the reflected camera rather than the real one. The predicted image after the motion therefore differs across the three hypotheses, and the gap between the predicted positions is exactly the separability $\Delta_{ij}(a)$ of equation 2 for the corresponding pair. Where every such gap stays below the perceptual threshold over the whole action set, the mechanism remains unidentifiable however capable the predictor happens to be. The slab hypothesis H_T follows the same logic, with a small axial displacement fixed by thickness and refractive index, and the figure omits it for clarity.

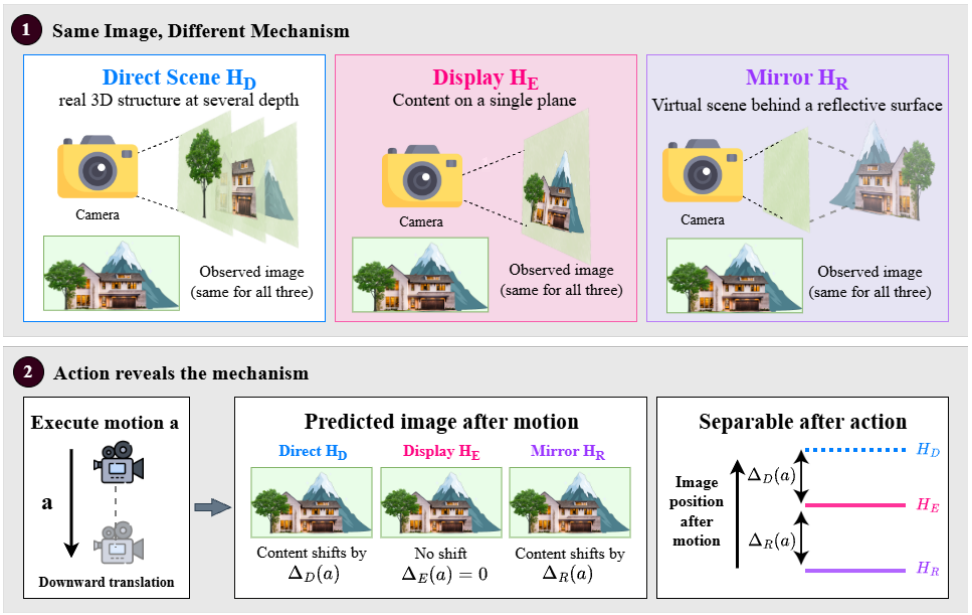


Figure 6: Three physical mechanisms are shown to produce the same observed image and are distinguished through controlled motion. (1) The same observation is produced by a direct 3D scene, a planar display, or a mirror-induced virtual scene. (2) When motion is applied, distinct image changes are produced by each mechanism, allowing them to be distinguished.

A.2 The Decision Pipeline

Figure 7 demonstrates how the pieces of Section 2 compose into a single decision. The system first observes the reference image I_0 and then carries the four candidate mechanisms H_D, H_E, H_R and H_T, which agree at that view. For every action in the bounded set, the transition $W(F_t, H_k, a)$ of equation 1 predicts the observation that each hypothesis implies, which turns one image into a table of predicted observations. Pairwise comparison over that table gives the separability $\Delta_{ij}(a)$ of each pair under each action. The decision stage then evaluates $\max_a \min_{i<j} \Delta_{ij}(a)$ and tests the resulting identifiability against the perceptual threshold ϵ. When $\mathcal{I}_\mathcal{A} \geq \epsilon$, the system executes the chosen action and then answers, and when $\mathcal{I}_\mathcal{A} < \epsilon$ it abstains before the camera moves.

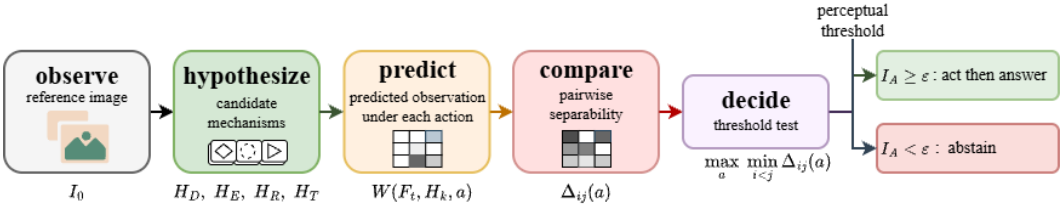


Figure 7: The pieces compose into a single decision. The system acts only when identifiability clears the perceptual threshold, and declines to answer otherwise.

A.3 The Real-Photograph Protocol in Detail

Figure 8 presents the construction of the real-data measurement and the place where we fit each quantity. A physical sheet or panel is flat, and reference disparity inside the annotated region should therefore stay planar. Where planarity fails, the stereo reference has locked onto displayed content rather than onto the panel and cannot settle what the predictor reported. We retain an image only when a plane fitted to that disparity attains $R^2 \geq 0.90$. We report the number of images that this rule excludes. Comparing two warps rather than two depth maps carries a further advantage. A depth-space comparison, by contrast, would score every pixel where the hypotheses disagree geometrically, including pixels whose appearance cannot tell the two apart. A photometric comparison

scores the same region through the image that an observer actually receives, and a featureless patch therefore contributes nothing whichever hypothesis holds. The median over valid pixels then guards against the small fraction of positions where occlusion or a rectification residual leaves one warp undefined. Both warps meanwhile read from the same photograph, hence exposure, sensor noise, and lens shading enter each side of the difference identically and cancel. What remains, then, is the disagreement that the two geometries imply, expressed in the intensity units of the photograph that produced the pair.

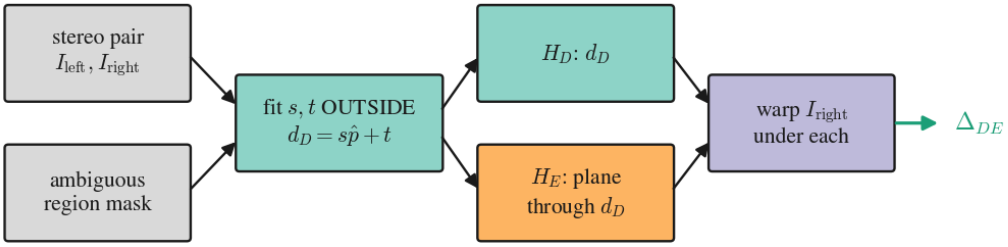


Figure 8: Construction of the real-data measurement, with the region where each quantity is fitted alongside the region where the same quantity is scored.

Table 4: The metrics reported throughout the paper. Here $\hat{H}$ denotes the predicted mechanism, $H^\star$ the true one, and $y_{\text{id}} = 0$ a case that no action resolves.

Metric	Definition	Question answered
CEA	$P(\hat{H} = H^\star)$	how often the mechanism is named correctly, abstention counted as an error
Committed acc.	$P(\hat{H} = H^\star \mid \text{answered})$	how often a given answer is correct
Abstention rate	$P(\text{abstain})$	how often an answer is withheld
FPCR	$P(\max_k p(H_k) > \tau \mid y_{\text{id}} = 0)$	how often a confident answer is given where none is warranted
Ident. AUROC	predicted against true resolvability	whether unresolvable cases are recognized
Normalized regret	chosen against best available action	whether a near-optimal intervention was selected
Motion cost	translation plus weighted rotation	how much movement the decision required

A.4 Minimum Causal Resolving Baseline

Proposition 1 (Minimum causal resolving baseline)**.** *Consider the pair (H_D, H_E) under a pure lateral translation of baseline B, with a pinhole camera of horizontal focal length f pixels, and $B \ll Z$. Two scene points at depths $Z_1 \neq Z_2$ then shift by approximately fB/Z_i, and they therefore produce a differential parallax $fB\,|1/Z_1 - 1/Z_2|$. A static planar display, by contrast, produces none once the screen-plane homography is compensated. Requiring that differential parallax to exceed the perceptual threshold δ yields the smallest resolving baseline*

$$B_{\min} > \frac{\delta}{f\,|1/Z_1 - 1/Z_2|}, \tag{13}$$

where δ coincides with ϵ whenever the motion term dominates the composite distance of equation 3.

The hypotheses of Proposition 1 are restrictive, and we check them before evaluating equation 13. A violated assumption therefore yields a missing prediction rather than a number. The proposition covers neither view-tracked displays, which no baseline resolves, nor mirrors, whose virtual scene carries genuine parallax.

A.5 Metric Definitions

Table 4 defines the metrics used throughout this paper. The unresolvable configurations that Section 3.1 retains take two forms, namely view-tracked displays observed within their aperture, which no baseline resolves, and mirrors arranged to mimic an opening of the same shape. Retaining both forms keeps the abstention metrics meaningful. A benchmark of resolvable cases alone would reward a method that answers everywhere, because every answer would then be available to some action.

B The Synthetic Benchmark

B.1 Choosing the Intervention, in Full

Figure 2c plots the mirror comparison quoted in Section 3.2, and Table 5 reports every objective on both studies. The selector study isolates the choice of action on reflection scenes alone, and the four-mechanism study repeats the comparison over the whole benchmark. Mirrors separate the two objectives because a mirror shares its parallax signature with a physically present scene over a wide range of motions, and only a narrow band of actions produces the visibility change that decides the pair. A summed objective spreads its weight over all six pairs and therefore drifts toward actions that separate easy pairs, which leaves 41.3% of its mirror commitments incorrect. Maximin selection, by contrast, optimizes the weakest pair by construction and resolves every mirror scene in every seed. The two objectives agree once the mirror pair leaves contention, which is why the four-mechanism columns differ by 0.014 while the mirror column differs by 0.413.

Table 5: Every selection objective on both studies, with eleven seeds each and chance at 0.250 over the four mechanisms. The selector study covers reflection scenes alone.

	Mirror committed		**Phase 2, four mechanisms**		
Objective	selector study	Phase 2	committed	abstains	forced CEA
Single-view classifier	—	—	$0.250{\pm}0.000$	0%	$0.250{\pm}0.000$
Passive multi-view	—	—	$0.510{\pm}0.020$	0%	$0.510{\pm}0.020$
Max-baseline	—	$0.925{\pm}0.048$	$0.981{\pm}0.011$	73%	—
Summed equation 6	$0.587{\pm}0.164$	$0.956{\pm}0.035$	$0.986{\pm}0.010$	69%	—
Maximin equation 7	$\mathbf{1.000{\pm}0.000}$	$\mathbf{1.000{\pm}0.000}$	$\mathbf{1.000{\pm}0.000}$	72%	$\mathbf{0.808{\pm}0.023}$

B.2 Calibrated Deferral and Imperfect Execution

Table 6 reports the effect of executing a chosen action imperfectly on the synthetic benchmark, and Table 7 reports split-conformal deferral on the real benchmark. A commanded motion rarely arrives exactly, and a selector that depends on the exact motion would lose its advantage in deployment. Under ± 1 cm and $\pm 0.5^\circ$ of execution error the hypothesis-blind proxy falls by 48.4% and a random choice by 29.0%, whereas Intervene3D falls by 2.9%. The margin explains the difference, because maximin selection prefers an action whose weakest pair separates by a wide margin, and a small perturbation of such an action still clears the threshold. The passive baseline anchors the comparison from the other side. A method that executes no action has nothing to perturb, hence its accuracy holds at 0.621 under both conditions and its change registers as exactly zero. Every acting method falls between that anchor and the proxy, and the ordering of the drops follows how tightly each one depends on the exact motion commanded. The paired t statistics separate the two regimes plainly, at -14.05 for the proxy against -3.05 for Intervene3D over the same eleven seeds. Table 7 then turns to calibration on real photographs. Both signals meet the coverage guarantee at all four target levels, which confirms that the conformal wrapper behaves as advertised in each case. Calibration, however, preserves the ranking of whichever score enters it, and the two scores rank oppositely at 0.344 for confidence against 0.645 for identifiability. Deferring on identifiability therefore lowers the failure rate among retained images by up to 12.1 points, while deferring on confidence raises that rate. The split of labour between wrapper and score explains the pattern. A coverage guarantee constrains how many predictions a system keeps, and the identity of those predictions follows from the score alone. Both rows of Table 7 consequently meet their target at every level while moving risk

in opposite directions. Deferring on confidence retains a set that fails at 76.6% against a base rate of 65.2%, which is worse than answering everywhere. Deferring on identifiability retains a set that fails at 53.1% at its best level. The wrapper contributes the guarantee, and the score contributes the ranking that decides whether the retained set improves at all. A practitioner reading coverage alone would accept both columns as sound, and the failure rate among retained images is what separates them. The same reasoning applies to any wrapper built on a fixed score, which is why Section 4 treats deferral rules and conformal prediction together.

Table 6: CEA under clean and perturbed execution, at ± 1 cm and $\pm 0.5°$ of error, with the paired t statistic over the same eleven seeds.

Method	Clean CEA	Perturbed CEA	Change	t
Passive multi-view (does not act)	0.621	0.621	$+0.0\%$	—
Generic next-best-view proxy	0.371	0.192	-48.4%	-14.05
Random intervention	0.409	0.291	-29.0%	-10.64
Max-baseline	0.502	0.478	-4.7%	-3.68
Intervene3D (summed)	0.560	0.544	$\mathbf{-2.9\%}$	-3.05

Table 7: Split-conformal deferral over 296 images at four target levels α, with the failure rate among retained images for each deferral signal.

		Failure rate among retained images					
Deferral signal	**AUROC**	α=0.2	α=0.3	α=0.4	α=0.5	**Coverage**	**Best gain**
No abstention	—			65.2%		—	—
Confidence, augmented	0.344	64.2%	65.8%	69.2%	76.6%	4/4	-1.0
Identifiability	**0.645**	58.1%	54.7%	**53.1%**	55.0%	4/4	$\mathbf{-12.1}$

C Real Photographs

C.1 Per-Checkpoint Results

Table 8 lists the per-checkpoint values summarized in Section 3.3 for the released checkpoints (Yang et al., 2024; Lin et al., 2026; Bochkovskiy et al., 2025; Wang et al., 2025; Kong et al., 2026; Yin et al., 2023; Piccinelli et al., 2024; 2026), together with parameter counts and family membership, while Figure 3 of the main text places all thirteen checkpoints on one photograph of a poster. Parameter count carries no signal here. The smallest checkpoint in the roster holds 24.8M parameters and the largest 952.0M, a range of $38\times$, and the fooled rates at the two ends sit within a point of each other. Sorting the column instead places DA3-Large at the top with 78.4% and the transparency fine-tuned variant at the bottom with 59.5%, two checkpoints from one lineage that differ in scale and data. Confidence AUROC stays below chance for all thirteen without a single exception, and identifiability stays above confidence for all thirteen, which is the pattern that Section 3.3 summarizes in aggregate.

C.2 Extended Results by Architecture Family

Table 9 averages within architecture family, which confirms that a single lineage does not drive the pattern. Mean fooled rates run from 61.8% for MoGe-3 to 74.3% for Depth-Anything-3, a spread of twelve points across seven families that differ in backbone, supervision, and training data. Mean confidence AUROC likewise stays below chance for every family, from 0.295 for UniDepth to 0.409 for Depth-Anything-3, while mean identifiability stays above chance throughout, from 0.522 to 0.692. The newest lineage carries the highest failure rate, which indicates that scale and recency alone move the phenomenon in the wrong direction rather than the right one. Family membership also explains the one visible dip. Depth-Anything-V2 averages 64.6% over four checkpoints, and one of those four is the transparency fine-tuned variant that fails least often across the whole roster. Removing that variant would raise the family mean rather than lower it, which places the dip in a training recipe instead of an architecture. The ordering between the two signals, moreover, holds inside every family without exception. Mean confidence stays below 0.5 in all seven, mean identifiability stays above 0.5 in all seven, and no family reverses the gap.

Table 8: Per-checkpoint values for the thirteen released checkpoints, with parameter counts and family membership. Confidence takes the better of raw confidence and a four-member ensemble.

Checkpoint	Family	Parameters	Fooled	Confidence AUROC	Identifiability AUROC
DA-V2-Small	Depth-Anything-V2	24.8M	65.2%	0.344	0.645
DA-V2-Base	Depth-Anything-V2	97.5M	67.2%	0.408	0.588
DA-V2-Large	Depth-Anything-V2	335.3M	66.6%	0.367	0.467
DA-transparency	Depth-Anything-V2	24.8M	**59.5%**	**0.234**	0.719
DA3-Small	Depth-Anything-3	24.8M	76.4%	0.465	0.479
DA3-Base	Depth-Anything-3	97.5M	68.2%	0.362	0.576
DA3-Large	Depth-Anything-3	410.9M	**78.4%**	0.399	0.512
DepthPro	DepthPro	952.0M	65.2%	0.328	0.659
MoGe-2	MoGe-2	330.9M	62.2%	0.334	0.642
MoGe-3-ViT-L	MoGe-3	330.9M	61.8%	0.360	0.595
Metric3D-Small	Metric3D	37.5M	64.5%	0.323	0.593
Metric3D-Large	Metric3D	343.0M	67.6%	0.406	0.593
UniDepth-v2	UniDepth	353.8M	66.2%	0.295	0.692

Table 9: The same thirteen checkpoints averaged within architecture family, with the number of checkpoints contributing to each row.

Family	Checkpoints	Mean fooled	Mean confidence	Mean identifiability
Depth-Anything-V2	4	64.6%	0.338	0.605
Depth-Anything-3	3	74.3%	0.409	0.522
DepthPro	1	65.2%	0.328	0.659
MoGe-2	1	62.2%	0.334	0.642
MoGe-3	1	61.8%	0.360	0.595
Metric3D	2	66.0%	0.365	0.593
UniDepth	1	66.2%	0.295	0.692
All	**13**	**66.8%**	**0.356**	**0.597**

C.3 Choosing Where to Look

Table 10 reports the full comparison. Fifty-five scenes without video content offer three or more camera positions of the same static content. We exclude video scenes because their frames differ by displayed content rather than by observer position, and treating those frames as an action set would count a change of stimulus as a change of viewpoint. The four rules consult increasing amounts of evidence, and the error inside the ambiguous region follows that ordering. Holding the first frame, at one extreme, consults nothing about the scene and leaves the error at 0.0671. A uniformly random choice consults nothing either, yet averaging over positions removes the worst of them and reaches 0.0591. The confidence of the predictor then consults the reference image alone and reaches 0.0439. The best-explained observation, finally, consults the view that each candidate action actually produces and reaches 0.0406, an improvement of 31.3% over the random choice and of 7.5% over confidence. Scoring inside the annotated region alone keeps the comparison independent of every other part of the image, where the four rules would otherwise agree.

C.4 Second Dataset with Ordinal Supervision

The second dataset provides ordinal relations between point pairs rather than dense depth, together with a per-point layer index authored independently of the present work. The ordinal form suits the question here, because a relation between two points holds under the scale and shift ambiguity that a monocular predictor leaves open. Splitting the relations by layer index then separates pairs that lie on a first visible surface from pairs that lie behind glass, with no quantity of ours entering the split. Both checkpoints answer the single-layer pairs well and lose ground on the multi-layer pairs, at 0.848 against 0.737 for DA-V2-Small and 0.843 against 0.669 for DepthPro. An annotation authored by another group therefore reproduces the ordering that Section 3.3 measures with stereo. The size of the gap tracks the architecture rather than the benchmark. DA-V2-Small loses 0.111

Table 10: Four rules for choosing a viewpoint, over 55 scenes and thirteen checkpoints, scored by error inside the ambiguous region.

Viewpoint chosen by	Error inside region	Fooled
First frame, no choice	0.0671	48.7%
Uniformly at random	0.0591	48.0%
The predictor's own confidence	0.0439	46.4%
Best-explained observation equation 8	**0.0406**	**44.1%**

across layers and DepthPro loses 0.175, against first-surface accuracies of 0.847 and 0.842 that differ by half a point. Two networks that agree closely on the surfaces they can see therefore diverge on the surfaces behind an interface, which is where the ambiguity under study lives.

Table 11: LayeredDepth validation (Wen et al., 2025), with ordinal accuracy split by the per-point layer index into single-layer and multi-layer point pairs.

Checkpoint	Single-layer	Multi-layer	Gap	First surface	Through glass
DA-V2-Small	0.848	0.737	**+0.111**	0.847	0.784
DepthPro	0.843	0.669	**+0.175**	0.842	0.741

C.5 Real Mirrors, with Annotated Contact Geometry

Every real-photograph result in the main text concerns physical sheets and panels, because the real split of the illusion benchmark contains no photograph of a physical mirror. Reflection therefore appears there on synthetic data alone. The present section closes that gap on real photographs, using mirror annotations produced by another group.

Mirror3D (Tan et al., 2021) annotates real mirrors with an instance mask and a three-dimensional mirror plane, layered on NYUv2 (Silberman et al., 2012), which requires no data agreement. The annotated plane serves as the appropriate ground truth here precisely because a depth sensor does not. A structured-light sensor returns either no signal or unreliable depth on a mirror, and the human-placed plane therefore remains the only statement of where the first touchable surface lies. We apply the same discipline as in the stereo protocol, fitting scale and shift outside the annotated region against reference depth and applying them inside that region. The comparison then runs in inverse-depth space, because the predictors return relative inverse depth rather than meters. The released annotations carry a 5% center crop, and we shift the principal point to match.

Table 2 in the main text reports the outcome over 123 annotated frames for the same thirteen checkpoints that the display benchmark uses, which places the two results on one roster. Eleven of the thirteen lie between 62.6% and 79.7%, around a mean of 67.0%. Working in inverse-depth space keeps the comparison on the terms that the predictors themselves use. A monocular network reports relative inverse depth, and converting to metres would require the very scale that the annotation leaves open on a mirror. Fitting scale and shift outside the annotated region against reference depth, then applying both inside, puts every checkpoint on the annotated plane's own footing without granting any of them information from the region under test. Figure 9 presents individual photographs from this set.

Two checkpoints stand apart at 41.5%. The first is the transparency fine-tuned variant, where training that targets one non-Lambertian mechanism appears to transfer to another. The same checkpoint also fails least often on the display benchmark, at 59.5%. The second is Metric3D-Large, whose small counterpart fails on 78.9% of the same frames, while the two agree closely on displays at 64.5% and 67.6%. The divergence therefore describes that checkpoint on this data rather than the family, and we leave the difference unexplained.

The magnitude of the failure matters alongside its frequency. Relative error inside the mirror region reaches 5.47× the error outside for Metric3D-Small, and two thirds of that checkpoint's frames place the first surface well behind the glass. A predictor that merely smoothed over the mirror, by contrast, would leave the error ratio near unity and the surface at the wall. Placing the surface metres behind the wall instead reproduces the reflected room, which is the same failure that a display produces

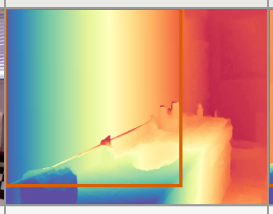
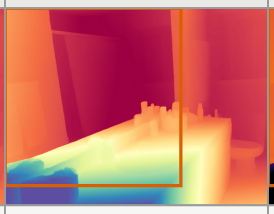
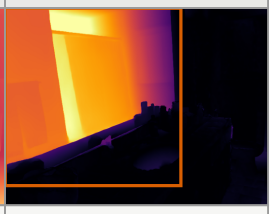
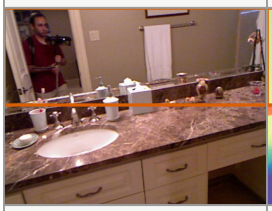
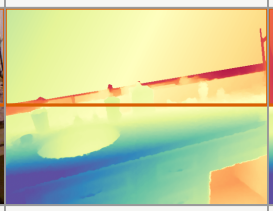
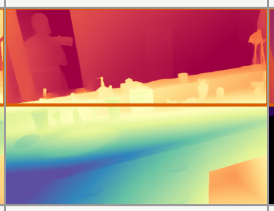
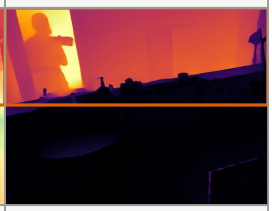
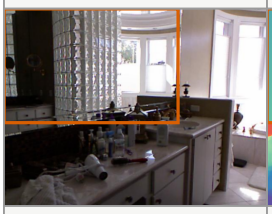
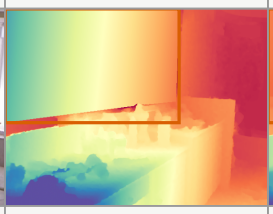
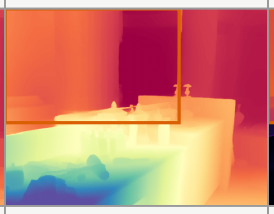
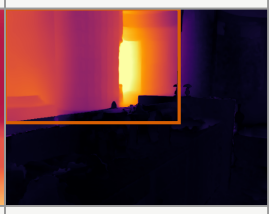
photograph	annotated plane	model reports	relative error
			
NYU 739	plane 1.4 m	reported 3.3 m	2.4× too far
			
NYU 664	plane 1.4 m	reported 3.3 m	2.3× too far
			
NYU 719	plane 1.6 m	reported 3.2 m	2.1× too far

Figure 9: Individual photographs from the annotated mirror set, with the human-placed mirror plane outlined in every panel.

with a depicted scene. The two mechanisms therefore differ in their optics while agreeing in what they do to a predictor. NYUv2 provides no camera poses and only accelerometer orientation, which leaves the action set empty. The experiment therefore carries a narrower message, namely that the phenomenon formalized here extends past displays and persists when an independent annotation, rather than a plane fit, states where the physical surface lies. Agreement between the two sources also rules out a family of alternative explanations. The display benchmark measures against a stereo reference, on photographs of printed sheets, posters, and monitors, annotated by one group. The mirror benchmark, by contrast, measures against a human-placed plane, on photographs of household mirrors, annotated by another group years earlier and for another purpose. A matching rate near two images in three consequently leaves little room for an artifact of one annotation protocol, one capture rig, or one class of scene content.

D Learned Abstention

D.1 Learned Abstention

We pool the features over the ambiguous region as three contrasts, namely the difference between the pooled representation inside and outside the region, the spread within the region, and the difference between a one-token boundary ring and its surroundings. The mechanisms separate at that ring. Contrasts serve as inputs because a representation encoding ambiguity should differ from its surroundings. We place tokens on the grid that the processor produces, and we choose the layer on inner folds. We then fit three classifier families, namely L_2-regularized logistic regression, histogram gradient boosting, and XGBoost, which keeps any conclusion independent of one inductive bias. The boosted models stay at depth 3 with at least 40 samples per leaf, because a deeper model would memorize a few thousand rows drawn from fewer than a hundred scenes. We also verify the depth convention of each checkpoint against reference stereo disparity before use, because the

attribute conventionally named `predicted_depth` carries metric depth in some families and inverse depth in others. An inverted convention would otherwise exchange near for far while still returning numbers of a plausible magnitude.

Table 12 presents the summary referenced from Section 3.6, and Figure 10 plots the transfer result. The next subsection then expands the leave-one-model-out column to every one of the thirteen held-out checkpoints.

Table 12: Four policies under grouped, nested, and leave-one-model-out protocols. Folds follow base scenes, which keeps every frame of one scene inside a single partition.

Policy	Grouped cross-validation	Nested cross-validation	Leave-one-model-out
all 13 checkpoints, 3,848 image–model pairs, 68 scenes			
Strongest single feature	0.766	—	0.768
Logistic regression	0.755	0.757	0.779
XGBOOST	—	0.834	—
Histogram gradient boosting	**0.834**	**0.831**	**0.881** (13/13)
single checkpoint, identical data for all three			
Strongest single feature	0.819	—	—
Hand-designed features	0.858	—	—
Frozen-encoder head	**0.887**	—	—

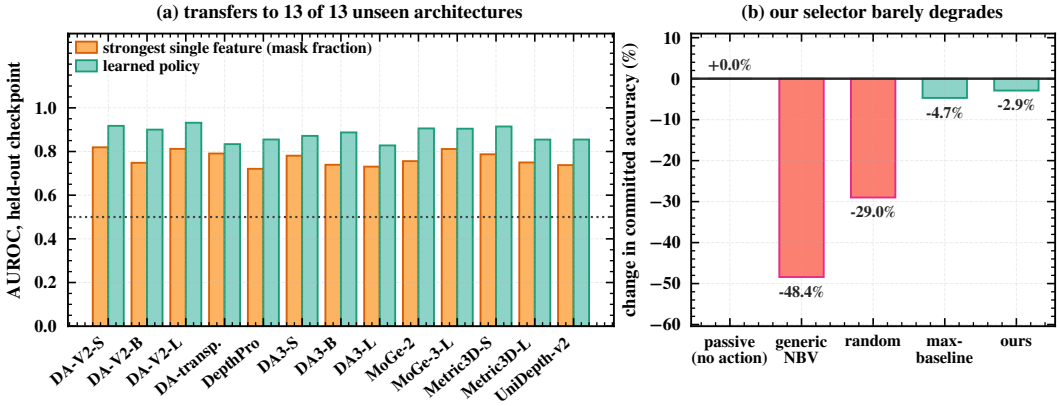


Figure 10: (a) Held-out checkpoints against the strongest single feature. (b) Committed accuracy under imperfect execution.

D.2 FULL LEAVE-ONE-MODEL-OUT RESULTS

Table 13 expands the leave-one-model-out protocol of Table 12 to every held-out checkpoint, together with the regularization sweep for the frozen-encoder head. Each row trains on twelve checkpoints and scores on the thirteenth, which leaves the held-out architecture unseen at fit time. The learned policy leads on all thirteen, by margins from $+0.043$ on the transparency fine-tuned variant to $+0.152$ on DA-V2-Base, for a mean margin of $+0.113$. A policy that had memorized one network would lose its advantage on an unseen one, and the uniform lead therefore points to a property of the image rather than of a particular predictor. The lower block sweeps the regularization strength across four decades, where the AUROC moves by 0.036 in total and decreases monotonically as the penalty weakens. The two hardest rows locate the residual difficulty. DA-transparency, on one side, yields the smallest margin at $+0.043$, while DA3-Large yields the lowest absolute value at 0.828. Both are checkpoints whose own failure rate departs furthest from the roster mean, at 59.5% and 78.4% against 66.8%, hence the policy meets an unfamiliar base rate when either one is held out. Predicting failure for a network that fails far more or far less often than its twelve neighbours remains the harder case, and a margin of $+0.043$ in the worst of them still favors the learned policy.

Table 13: The leave-one-model-out protocol over every held-out checkpoint. The lower block sweeps the regularization strength C for the frozen-encoder head.

Held out	Learned	Baseline	Δ	Held out	Learned	Baseline	Δ
DA-V2-Small	0.917	0.819	+0.098	DepthPro	0.855	0.721	+0.134
DA-V2-Base	0.900	0.748	+0.152	MoGe-2	0.906	0.756	+0.150
DA-V2-Large	0.932	0.812	+0.120	MoGe-3-ViT-L	0.904	0.811	+0.093
DA-transparency	0.834	0.791	+0.043	Metric3D-Small	0.915	0.787	+0.127
DA3-Small	0.871	0.781	+0.091	Metric3D-Large	0.855	0.750	+0.105
DA3-Base	0.887	0.739	+0.148	UniDepth-v2	0.855	0.738	+0.117
DA3-Large	0.828	0.731	+0.097				
mean 0.881 against 0.768, margin +0.113, ahead on 13/13							
frozen-encoder head: AUROC against regularization strength C, grouped five-fold, one checkpoint							
C	0.001	0.005	0.020	0.100	0.500	2.000	10.000
AUROC	**0.879**	0.875	0.873	0.866	0.859	0.853	0.843

D.3 Full Abstention Retention Curve

The main text quotes three operating points, and Table 14 reports the complete curve. Every threshold comes from out-of-fold scores under folds split by scene, which keeps the choice of operating point independent of the images that it scores. The failure rate then falls monotonically from 65.2% at full coverage to 23.6% at 30% coverage, with no reversal along the way. In practical terms, answering seven inputs in ten avoids 89 of the failures, and answering half avoids 137. Monotonicity matters for deployment, because an operator who tightens the gate can rely on the failure rate moving in the intended direction rather than on a single tuned point.

Table 14: The frozen-encoder gate on 296 images from DA-V2-Small, at coverage levels from 100% down to 30%, against a base failure rate of 65.2%.

Inputs answered	Failure rate among them	Change	Failures avoided
100%	65.2%	—	0
90%	61.3%	−3.9	30
80%	56.5%	−8.7	59
70%	50.2%	−15.0	89
60%	43.3%	−21.9	116
50%	37.8%	−27.4	137
40%	29.7%	−35.5	158
30%	23.6%	−41.6	172

E Robustness and Controls

E.1 Threshold Sensitivity

Three thresholds carry the decisions in this work, and the present section reports their influence. Sweeping the perceptual threshold from 0.25 to 3.0 px leaves every reported quantity unchanged, at 34.3% abstentions, a CEA of 0.560, a committed accuracy of 0.855, and an FPCR of 0.000, and only at 5.0 px does the abstention rate move to 37.8%. Table 15 sweeps the admissibility criterion, and Table 16 reports what that criterion removes by category. We recompute every number below from the artifacts of the original runs, without re-executing an experiment or redrawing a seed. Each row therefore records the decision that the same system would have taken under a different threshold on identical data. The admissibility sweep answers a question about selection rather than about accuracy. A criterion that admitted only the images most favorable to the measure would inflate every number downstream of it. Moving the criterion from $R^2 \geq 0.70$ to $R^2 \geq 0.98$, however, changes the retained fraction from 70.5% to 55.4%, a swing of fifteen points in how much of the benchmark remains. The fooled rate over that swing meanwhile moves by 1.6 points and the identifiability AUROC by 0.051, both in the direction that a stricter reference measurement predicts. Neither quantity, therefore, depends on where the criterion falls within the range that we tested.

Table 16 then reports what the criterion removes, category by category. Exclusions concentrate almost entirely where a monitor fills the frame, at 99 of 99 images, and thin elsewhere, at 49 of

Table 15: The admissibility criterion swept from $R^2 \geq 0.70$ to 0.98 on DA-V2-Small, with the retained fraction and two AUROC values.

R^2 threshold	**retained**	**fooled**	**identifiability AUROC**	**region-size AUROC**
0.70	70.5%	64.8%	0.641	0.813
0.80	69.5%	65.2%	0.637	0.820
0.85	67.7%	65.9%	0.637	0.822
0.90	**65.1%**	**65.2%**	**0.645**	0.819
0.95	63.1%	64.8%	0.656	0.819
0.98	55.4%	64.3%	0.692	0.827

Table 16: What the admissibility criterion retains and removes in each category, with median reference planarity for both groups.

Category	**total**	**retained**	**excluded**	**planarity retained**	**planarity excluded**
Printed sheet on a table	104	97	7	0.996	0.816
Poster beside an object	153	104	49	0.996	0.547
Monitor in the frame	99	95	4	0.995	0.880
Monitor filling the frame	99	0	99	—	**0.015**

153 posters, 7 of 104 sheets, 4 of 99 monitors. Median reference planarity, in turn, separates the two groups cleanly, at 0.995 or above among retained images against 0.015 where a monitor fills the frame. A displayed scene that occupies the whole frame leaves the stereo matcher no physical structure to lock onto, and the resulting disparity describes the displayed content instead of the panel. Such a measurement cannot arbitrate between the two hypotheses under test, and the criterion removes it for that reason alone.

Computational Cost: The decision itself remains inexpensive. On the four-mechanism benchmark the selector enumerates $|\mathcal{A}| = 33$ actions against six hypothesis pairs, which amounts to 198 separability evaluations. The full enumeration then takes 0.079 ms per scene on a central processing unit in NumPy, without graphics acceleration. On real photographs the depth forward pass dominates the cost rather than the decision, at 462 ms per image end to end for the smallest checkpoint and 1140 ms for the largest. Scaling to a larger action set, in turn, grows linearly in $|\mathcal{A}|$. The evaluations also remain independent, which allows a direct parallelization of the enumeration over available cores.

Conformal Calibration on the Identifiability Score: The split-conformal results of Table 7 already use identifiability as the nonconformity score. The coverage guarantee therefore arrives together with the reduction in selective risk, since the procedure meets that guarantee at all four levels while lowering the failure rate among retained images at every level.

E.2 Sensitivity to the Occlusion Weight

The composite distance of equation 3 weights an occlusion term against a motion term, with $\lambda_o = 4.0$ throughout. A single constant that decides which action the selector chooses raises the question of how far the conclusions depend on that constant. We therefore rerun the same eleven seeds at $\lambda_o \in \{1, 2, 4, 8, 16\}$, a $16\times$ range, with every other setting fixed. Table 17 reports the outcome. Committed accuracy moves by 0.018 across the whole range, and the selector stays ahead of the passive multi-view baseline at every weight. The reflection column follows the summed objective of Table 1 and holds its rate throughout. Two larger weights edge past the value used in the paper, and the conclusions rest on the ordering alone.

E.3 Threshold, Temperature and Sensor Scale

Three constants stay fixed throughout and receive no tuning, namely the perceptual threshold $\epsilon = 1.0$ px, the inverse temperature $\beta = 1.0$ of the belief update, and the commitment threshold $\tau =$

Table 17: Committed accuracy over eleven seeds at five occlusion weights λ_o, with the value used in the paper in bold and reflection scenes at right.

λ_o	Intervene3D (sum)	Passive multi-view	Max-baseline	Reflection only
1	0.841 ± 0.051	0.621 ± 0.043	0.818 ± 0.020	0.453 ± 0.173
2	0.853 ± 0.051	0.621 ± 0.043	0.822 ± 0.019	0.580 ± 0.162
4	$\mathbf{0.855 \pm 0.051}$	0.621 ± 0.043	0.824 ± 0.019	0.587 ± 0.164
8	0.857 ± 0.052	0.621 ± 0.043	0.824 ± 0.019	0.593 ± 0.164
16	0.859 ± 0.055	0.621 ± 0.043	0.824 ± 0.019	0.598 ± 0.173

0.45. A threshold stated in pixels depends on the sensor for its meaning, and we therefore sweep resolution alongside the other two constants. Table 18 reports all three over the same eleven seeds.

The constants β and τ enter at evaluation, which lets those arms share one dataset under a paired comparison. Resolution instead changes the rendered scene, and each resolution arm regenerates its data under a distinct name, which also reseeds the draw. Those arms accordingly admit only an unpaired comparison, and differences of a few thousandths there include sampling noise.

Table 18: Committed accuracy over eleven seeds for the belief temperature β, the commitment threshold τ, and sensor resolution. The resolution block admits an unpaired comparison alone.

Configuration	Intervene3D (sum)	Max-baseline	Passive multi-view	Reflection only
belief temperature β, paired				
0.25	0.838 ± 0.053	0.814 ± 0.021	0.621 ± 0.043	0.450
0.5	0.854 ± 0.051	0.821 ± 0.019	0.621 ± 0.043	0.580
1.0 (reported)	$\mathbf{0.855 \pm 0.051}$	0.824 ± 0.019	0.621 ± 0.043	0.587
2.0	0.855 ± 0.051	0.824 ± 0.018	0.621 ± 0.043	0.587
4.0	0.855 ± 0.051	0.824 ± 0.019	0.621 ± 0.043	0.587
commitment threshold τ, paired				
0.30	0.855 ± 0.051	0.824 ± 0.019	0.621 ± 0.043	0.587
0.45 (reported)	$\mathbf{0.855 \pm 0.051}$	0.824 ± 0.019	0.621 ± 0.043	0.587
0.60	0.855 ± 0.051	0.824 ± 0.019	0.621 ± 0.043	0.587
0.80	0.855 ± 0.051	0.824 ± 0.019	0.621 ± 0.043	0.587
sensor resolution and ϵ, unpaired				
160×120, $\epsilon = 1.0$ px	0.864 ± 0.027	0.820 ± 0.024	0.620 ± 0.044	0.619
160×120, $\epsilon = 0.5$ px	0.864 ± 0.027	0.820 ± 0.024	0.620 ± 0.044	0.619
320×240, $\epsilon = 1.0$ px (used)	$\mathbf{0.863 \pm 0.046}$	0.821 ± 0.047	0.605 ± 0.038	0.626
640×480, $\epsilon = 1.0$ px	0.858 ± 0.043	0.834 ± 0.028	0.606 ± 0.061	0.599
640×480, $\epsilon = 2.0$ px	0.858 ± 0.043	0.834 ± 0.028	0.606 ± 0.061	0.599

Committed accuracy moves by 0.017 across a sixteenfold range of β, by nothing at all across τ, and by 0.006 across a fourfold range of linear resolution, which is a sixteenfold range in pixel count. Moreover, the selector stays ahead of the passive baseline in every configuration.

The β result forms a plateau, because at 0.5 and above the update already stays decisive enough to leave the posterior ordering unchanged. Only at 0.25 does the discount on the evidence cost accuracy. The invariance to τ has a structural cause. The identifiability rule of equation 5 declines first, and no unresolvable case therefore reaches the commitment threshold. The commitment threshold then governs only cases that the action set has already separated. A larger τ does change the internal confidence rate on identifiable scenes, from 1.000 at $\tau \leq 0.45$ to 0.789 at $\tau \geq 0.60$, while the commitments that the system makes remain the same throughout.

The resolution result answers the question that a threshold stated in pixels raises. Holding ϵ at 1.0 px while the sensor changes tests an absolute pixel threshold, and scaling ϵ with the resolution tests the same angular threshold. Both arrangements produce identical outcomes at every resolution, which follows from the bimodal separability on this benchmark. Almost no scene lies between 0.5 and 2.0 px, and any threshold inside that band picks out an identical set of scenes.

E.4 Comprehensive Analyses

We recompute every number in this section from the per-image records that the reported evaluations already produced, without re-running a model or redrawing a seed.

The Failure Label and Whether Region Size Is Doing the Work: A predictor counts as fooled when its relative error inside the annotated region exceeds its error outside that region. The strongest trivial baseline uses the fraction of the frame that the region occupies, at 0.819 against 0.645 for identifiability, which invites the reading that the measure detects large regions. Table 19 separates the two quantities by rescoring within terciles of region size. Once region size stays roughly fixed, that baseline falls from 0.819 to 0.542, close to chance and characteristic of a confound, whereas identifiability holds at 0.618. Under a label demanding a margin ($e_{\text{in}} > 1.5\, e_{\text{out}}$, base rate 51.7%) the two reach 0.601 and 0.772, which makes the ordering independent of the chosen inequality.

Table 19: Identifiability and the region-size baseline rescored within terciles of region size, over 296 admissible images.

Region size	Images	Identifiability	Region size
0.064–0.307	99	0.429	0.461
0.307–0.370	98	0.694	0.792
0.370–0.793	99	0.731	0.374
mean within strata	—	**0.618**	0.542
pooled	296	0.645	**0.819**

Whether Δ Describes the Scene or the Predictor: Both hypotheses come from the aligned disparity of the predictor itself, which raises the question of whether Δ measures a property of the scene or the habits of one network. The test proceeds directly, by scoring the Δ of one checkpoint against the failures of a different checkpoint. Over 13 checkpoints and 156 ordered cross pairs, the mean cross-model AUROC reaches 0.637, against 0.597 when each checkpoint faces its own failures. Every cross pair also stays above chance, over a range from 0.504 to 0.859. A separability computed from one predictor therefore anticipates the failures of another predictor at least as well as its own, which characterizes a quantity carrying scene structure rather than predictor idiosyncrasy.

What the Admissibility Criterion Removes: The criterion retains 296 of 455 images. The excluded group differs from the retained one almost entirely in the property that the criterion tests. Median reference planarity reaches 0.024 among excluded images against 0.995 among retained ones, which is the signature of stereo matching that locks onto displayed content rather than onto the panel. Median region size stays nearly identical across the two groups, at 0.333 against 0.357, and the filter has no preference for the size of the ambiguous region. Exclusions also concentrate in the video category, at 99 of 159 images, the category in which a monitor fills the frame.

Which Decisions the Perceptual Threshold Governs: Identifiability AUROC rests on ranks and therefore stays invariant to ϵ by construction, since the threshold moves which images receive an answer rather than how the score orders them. The threshold governs instead the split between answering and abstaining. Against a base failure rate of 65.2%, committing only where $\Delta \geq 0.5$ answers 59.1% of images at 53.1%, a reduction of 12.1 points. Committing where $\Delta \geq 1.0$ instead answers 34.8% at 59.2%. The split-conformal results of Table 7 report this regime, in which a useful fraction of the stream still receives an answer. In the extreme tail, below 10% answered, the ordering no longer holds, and the operating points reported here retain at least a third of the images.

A Flat-Panel Hypothesis Independent of the Predictor: Both hypotheses come from the aligned disparity of the predictor. We therefore built a second H_E with no access to the prediction, namely a plane fitted to reference disparity in an annulus outside the annotated region and extrapolated inward. The fit lies where the alignment already lies, and the reference stays outside the computed quantity. On identical images the prediction-based construction reaches 0.645 and the surroundings-based construction reaches 0.532, with a median Δ of 1.612 against 0.721, a factor of 2.23. A hypothesis matching the panel predicts observations close to H_D and therefore a small Δ. A value inflated

twofold, by contrast, reports a plane in the wrong place, which is the outcome for a poster beside an object or a monitor standing proud of the wall once assumed coplanar with its surroundings. The measure needs an accurate flat-panel hypothesis, and the aligned prediction yields a closer plane than the surrounding disparity. The cross-model result above indicates that the dependence rests on the hypothesis rather than on one network, since Δ from one checkpoint predicts the failures of another at 0.637.

Whether Cheap Spatial Cues Would Improve the Learned Gate: Region size, local texture, reference planarity, and the disparity gap cost nothing to compute, which raises the question of whether adding them lifts the frozen-encoder gate. Table 20 reports three arms on identical folds split by base scene, at the same encoder layer, and the arms differ only in their features. The cues alone reach 0.799, the pooled encoder features reach 0.873, and the two together reach 0.873. Adding the cues changes nothing, which indicates that the representation computed by the encoder already carries whatever those cues measure. Choosing the layer on inner folds rather than fixing the layer at the last one raises the gate to 0.887, and the selected layers remain consistently early ones.

Table 20: Three feature sets scored out of fold on identical folds, over 296 images carrying 193 failures.

Features	Dimension	Out-of-fold AUROC
Spatial cues alone	4	0.799
Frozen encoder features	1152	**0.873**
Encoder features + cues	1156	0.873

Where the Region Boundary Is Drawn: The annotated regions come from the benchmark authors, which raises the question of whether the conclusions depend on a boundary placed a few pixels differently. Table 21 grows and shrinks every region by a fixed radius and re-runs the evaluation. Across a 16px range the identifiability AUROC moves between 0.621 and 0.653, against 0.645 as reported. Eroding by 8 px admits more images, at 395 against 296, and lowers the fooled rate to 50.6%. The smaller region explains that direction, because a narrower region stays planar more often in the reference measurement and also drops the boundary pixels where the mechanisms disagree most.

Table 21: Every annotated region grown or shrunk by a fixed radius on DA-V2-Small, with the evaluation re-run at each radius.

Boundary shift	Admissible	Fooled	Identifiability	Median region size
-8 px	395	50.6%	0.621	31.6%
-4 px	308	64.9%	0.653	35.3%
0 (as reported)	**296**	**65.2%**	**0.645**	35.7%
$+4$ px	290	64.1%	0.649	36.0%
$+8$ px	285	63.5%	0.649	36.5%

A Robust Fit for the Admissibility Criterion: An ordinary least-squares plane fitted to reference disparity decides admissibility, and a handful of outlying measurements can drag such a fit. We therefore recomputed the criterion with a random sample consensus fit that ignores those measurements. On all 455 images the two criteria select the same set at every threshold tested, agreeing on 100% of images at $R^2 \geq 0.80$, 0.90, and 0.95, and on 99.8% at 0.85. A few poor disparities pulling the plane thus do not produce the criterion.

Whether the Advantage Depends on Being Allowed to Rotate: The action set contains rotations of $\{5^\circ, 10^\circ\}$ in yaw and pitch alongside the translations, which raises the question of how much of the advantage of the selector follows from permission to rotate. Removing the rotations entirely reduces $|\mathcal{A}|$ from 33 to 25. Committed accuracy then stays at 0.855 ± 0.016, against 0.848 ± 0.044 for a matched control run at the full action set over the same eleven seeds. We quote that control rather than Table 1, because this ablation redraws the benchmark and the two arms therefore compare with each other rather than with the main run. The selector stays ahead of the passive multi-view baseline

at 0.629 by the same margin. The advantage thus rests on choosing which motion to execute, not on access to a richer motion vocabulary. The distinction matters for the intended deployment, where a device may translate a little without rotating freely.

The Two Selection Objectives on Every Mechanism: The argument for the weakest-pair criterion runs through mirrors, and the present paragraph reports the other three mechanisms as well. Over 11 seeds, committed accuracy under the summed objective reaches 1.000 on direct, emissive, and transmission scenes and 0.956 ± 0.035 on reflection, whereas maximin selection reaches 1.000 on all four. The maximin criterion accordingly leads or draws on four mechanisms out of four and separates from the summed objective on exactly one. Mirrors receive the emphasis because that case alone changes the outcome, exactly as equation 6 predicts. A sum reaches its maximum at an action separating pairs outside contention, and the mirror pair remains the pair that such an action can leave unseparated.

Which Layer the Gate Uses and How the Head Pools: The head pools the frozen tokens of the encoder over the annotated region and its surroundings. The pooling keeps three quantities, namely the difference between the pooled representation inside and outside the region, the spread within the region, and the difference between a one-token boundary ring and the surroundings. The boundary term enters because the mechanisms separate at the aperture rim rather than in the interior. We place tokens on the grid that the processor actually produces, which departs from square and varies with input aspect ratio. An assumed square grid would otherwise shear the feature map and misassign which tokens fall inside the region.

We choose the layer on inner folds alone, and no reported number comes from the split that made that choice. Across the five outer folds the selection lands on layers 3 to 5 of 13 and at no point on the last, which places the evidence available to the gate in early-to-mid features rather than in the final representation. A layer fixed at the last yields 0.873, whereas a selected layer yields 0.887. Sensitivity to the regularization strength remains mild, with the AUROC moving between 0.843 and 0.879 across four decades of C at a fixed layer.

What Each Signal Observes and Why: Figure 2d places a separability computed from a stereo pair beside a confidence computed from a single image. The two quantities carry different sensory budgets, and that difference states the claim. Under the premise of this work the reference image does not determine the physical explanation. A signal confined to that image therefore has nothing to draw on, because the same pixels stay consistent with every hypothesis, and no calibration recovers a distinction that those pixels do not carry. A monocular confidence succeeding here would contradict the premise rather than support Intervene3D. The evidence requires the observation that the action produces, which equation 4 formalizes and over which equation 5 operates.

We therefore report two families of comparison, which answer different questions. Against the reference image alone, the learned policies of Section 3.6 and the trivial baselines quoted beside every measure all take what the predictor computes from one view, and those comparisons stay like for like. Against the executed action, the comparison between separability and confidence establishes that the action provides something the single view does not, which is the point at issue. Reporting only the first family would leave the central claim untested. Reporting only the second would invite the reading that the advantage follows from the extra observation rather than from the use made of that observation, and both families therefore appear above.

Visibility and the Unused Distance Term: Z-buffering the predicted geometry of each hypothesis decides visibility under that hypothesis, and a landmark counted as occluded under one mechanism and visible under another therefore contributes to $D_{\text{occlusion}}$. At grazing incidence the decision follows the predicted geometry rather than a smoothed variant, which keeps the term discrete and lets a single unambiguous appearance or disappearance carry the weight described in Section 2.3. The feature term D_{feature} of equation 3 carries weight $\lambda_f = 0$ throughout and stays unevaluated. The definition retains that term because the composite distance covers a general hypothesis family, where a learned appearance distance would enter at the same position. The zero weight meanwhile keeps D expressed in pixels.

Landmarks, Belief and Runtime. Each synthetic scene carries 40 content landmarks placed by jittered stratified sampling over the interface extent, on a 7×7 grid over forty of its cells, together with 3 observer markers. The camera itself measures 320×240 with $f_x = f_y = 320$. Stratification serves coverage of the region, which determines whether a visibility change becomes observable at all. The belief update uses $\beta = 1.0$ and a posterior floor of 10^{-6}, which stops one unlucky observation from eliminating a hypothesis permanently. Cost stays small at both ends. Identifiability over $|\mathcal{A}| = 33$ takes 0.079 ms per scene, with each hypothesis predicting its consequence in closed form and no model running per action. The gate then costs 0.0029 ms per region on 1152-dimensional pooled features, on top of an encoder pass that the system already performs.

F Scope, Limitations and Reproducibility

F.1 Extending Beyond Planar Optics

The hypothesis family used here stays planar, covering a flat panel, a planar mirror, and a planar slab. Curved mirrors and non-planar transparent surfaces occur commonly, and the present section states precisely what would change to admit them and what would remain.

The Formulation Does Not Change: No part of the decision rule depends on planarity. Separability equation 2 holds between the predicted consequences of any two hypotheses under any action, identifiability equation 4 maximizes that quantity over the action set, and the abstention rule equation 5 minimizes over contending pairs against a perceptual threshold. A curved mirror, for instance, enters as a further H_k carrying its own transition W, and reaches the maximin selector equation 7 as one more pair without any change to the criterion. The belief update equation 9 behaves the same way, since that update requires a residual and nothing further. Adding a mechanism therefore changes the hypothesis set and the transition while leaving the deciding machinery untouched.

What Does Change Is the Guarantee and That Is the Real Barrier: The present construction derives its value from closed-form transitions, since a planar mirror reflects exactly across a plane and a planar slab displaces content axially by $d\,(1 - 1/n)$. Both branches remain exact. Reference views therefore match to a maximum deviation of order 10^{-14} px, and any divergence after an action follows from the mechanism alone. A curved surface, however, admits no exact transition, and two routes remain open, where each route weakens a different property:

- **A parametric surface**, a spherical mirror of given radius say, keeps a closed-form transition and therefore keeps the matched guarantee exactly. Such a family extends in a controlled way, although only to the surfaces that the parameterization covers, which replaces one restriction with a weaker restriction rather than removing the restriction altogether.
- **A learned transition** covers arbitrary geometry while forfeiting the property on which the design depends. Once a separate fit produces W per mechanism, a difference between two predictions may follow from a difference between fitted models instead of physical mechanisms, which is exactly the confound that the shared projection removes. The matched-counterfactual guarantee would then hold only approximately, by an amount that itself resists bounding.

The synthetic results state claims about the four named mechanisms, and the real-data results concern physical and emissive surfaces, which are planar in fact rather than by assumption. The second dataset used here contains real glass, where annotations authored independently confirm the same ambiguity (Table 11). The phenomenon therefore extends to real transparent surfaces, although we do not evaluate mechanism recovery on those surfaces. Curved optics remain outside the measurements, where the decision rule would apply unchanged while the benchmark guarantee behind the rule would lapse.

F.2 Scope and Limitations

Three limits bound what the reported numbers establish. First, the synthetic identifiability AUROC is a contrast rather than an absolute score. One computation produces both the identifiability prediction and the ground-truth label on the synthetic benchmark. The value of 1.000 in Table 1 therefore

stands only against the 0.505 and 0.748 that the classifiers reach on identical scenes. Second, abstention holds under encoder noise, with one seed apart. The clean configuration reaches a false physical certainty rate of 0.000 ± 0.000 across all 11 seeds, and the noisy-encoder ablation reaches 0.000 in ten of eleven seeds, averaging 0.008 ± 0.026. Third, real photographs provide a single action. A calibrated stereo pair yields $|\mathcal{A}| = 1$, hence equation 4 reduces to one term and the selection results of Table 5 rest on the synthetic benchmark. Evaluating the choice of intervention on real imagery calls for a multi-view capture or a controllable camera.

Two contemporaneous preprints appeared during the preparation of this work, and we record them here for completeness. Nguyen et al. (2026) probe hallucinated structure elicited by planar yet perceptually ambiguous inputs, contributing a benchmark with context variation, two second-order magnitude-based scores, and a training strategy that reduces the effect. The two studies agree on the phenomenon and differ in the response offered, namely a reduction in hallucination magnitude there and a decidability criterion with an explicit abstention here. The two accounts complement each other. Sun et al. (2026) forecast the temporal evolution of frozen geometry-foundation features rather than future pixels, which establishes that such features carry predictive structure and admit direct operations. Section 3.6 uses the same frozen substrate for a different purpose, namely decoding whether a scene remains decidable rather than forecasting how that scene will evolve. The agreement on that substrate therefore offers independent support for the read-out described there.

F.3 Reproducibility

We average all synthetic results over 11 seeds with the benchmark redrawn per seed, which makes the reported intervals a statement about the data-generating process. Splits follow base scenes throughout, in the synthetic benchmark and in the learned-policy protocols alike, which keeps variants or frames of a single scene inside one partition. We use every depth checkpoint exactly as released, and we train or tune nothing on any evaluation set. We verify the output convention of each checkpoint against reference stereo disparity before use, because the attribute conventionally named `predicted_depth` carries metric depth in some families and inverse depth in others. One checkpoint, UniDepth-v2, runs without its optimized attention kernel on the available hardware, and its numbers form a lower bound on the released model rather than a tuned comparison. We release the code upon acceptance for reproducibility and encourage further research on this topic.

F.4 Use of Generative Artificial Intelligence

We developed the research problem, conceptual framework, hypothesis family, methodological architecture, theoretical results, and benchmark design independently, before using any generative artificial intelligence tool. Such tools then served only as a general-purpose assistant under our direct supervision, for implementation and routine programming, plotting scripts, supplementary controls, dataset cleaning, and refinement of text that we had already drafted. We reviewed every assisted output before inclusion, verified all citations against the original sources, and obtained every reported number from experiments that we ran with the released code.