One Scene, Two Depths: Probing Geometric Ambiguity in Monocular Foundation Models

Xiaohao Xu[1], Feng Xue[1], Xiang Li[2], Haowei Li[1], Shusheng Yang[3], Tianyi Zhang[2], Matthew Johnson-Roberson[2,4], and Xiaonan Huang[1]

[1] University of Michigan, Ann Arbor, MI, USA
{xiaohaox, fengxe, haoweili, xiaonanh}@umich.edu
[2] Carnegie Mellon University, Pittsburgh, PA, USA
{xl6, tianyiz4}@andrew.cmu.edu,
[3] New York University, New York, NY, USA
shusheng.yang@nyu.edu
[4] Vanderbilt University, Nashville, TN, USA
matthew.johnson-roberson@vanderbilt.edu

GitHub Repo: https://github.com/Xiaohao-Xu/Ambiguity-in-Space

Abstract. A faithful 3D world representation should account for layered geometry, where a single camera ray may contain multiple visible and geometrically valid surfaces. Monocular depth estimation, however, reduces this structure to one scalar depth per pixel. Transparent scenes make this ambiguity measurable: the same ray can pass through foreground glass and observe the background, turning the supervised target into a convention of annotation, data, and training rather than a scene-intrinsic truth. A learned predictor exposes this convention as its depth-layer preference. We introduce `MultiDepth-3k` (`MD-3k`), a sparse two-layer ordinal benchmark for measuring depth-layer preference and multi-layer spatial relationship accuracy (ML-SRA). On `MD-3k`, leading depth foundation models exhibit diverse layer preferences under standard RGB input, showing that the same layered geometry can be resolved differently across models. We further find that Laplacian Visual Prompting (LVP), a training-free spectral input transformation, can substantially change the reported layer for certain frozen models. The strongest RGB/LVP pair, DAv2-L, reaches 75.5% ML-SRA. These results suggest that depth foundation models may express complementary geometric hypotheses that standard RGB inference leaves unexpressed. We invite the community to rethink depth supervision and evaluation through an ambiguity-aware lens, where multiple valid 3D interpretations are treated as geometric structure to be measured, preserved, and expressed.

Keywords: Monocular Depth Estimation · Multi-Layer Geometry · Visual Prompting · Foundation Model · Layered 3D Scene Understanding

1 Introduction

Depth maps are a compact interface between images and 3D reasoning: each pixel is assigned a distance and can be lifted into a point for reconstruction,

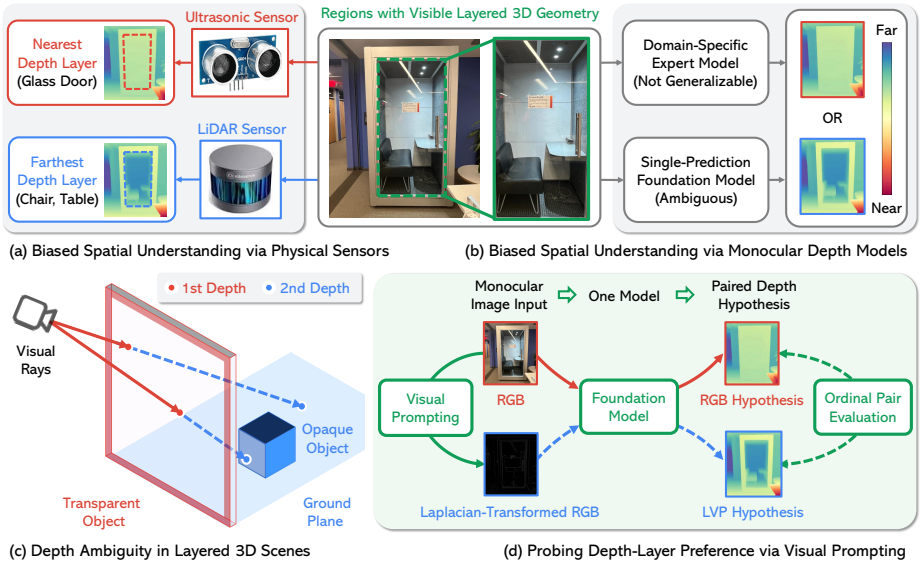


Fig. 1: Rethinking geometric ambiguity for 3D spatial understanding. (**a**) Ambiguous layered scenes can contain multiple visible surfaces along one ray, while single-depth supervision records only one biased layer. (**b**) This collapse turns multi-layer geometry into a dataset-shaped scalar target. (**c**) A single line of sight intersects multiple surfaces in transparent scenes. (**d**) Laplacian Visual Prompting (LVP) can surprisingly modulate the predicted layer of certain frozen models without retraining, revealing complementary ordinal behavior at the benchmark level.

navigation, and scene understanding. This interface quietly assumes that each visual ray has one surface to report. The assumption is usually acceptable in opaque scenes, but it becomes fragile when visibility is layered. Modern depth foundation models [24, 34, 35] inherit this interface: despite broad pretraining, they predominantly follow the **single-depth prediction** paradigm, mapping an image to one depth value per pixel. Transparent scenes expose the tension clearly: one ray can carry evidence from both a transparent foreground and the scene behind it (Fig. 1c). Collapsing this multi-layer geometry into one target turns the depth label into a *layer convention* shaped by annotation, dataset construction, and single-depth training, rather than a unique property of the image or scene.

This convention is further shaped by the mixture of supervision. Sensor-derived labels may emphasize different physical layers: ultrasound can return the proximal glass, while LiDAR may pass through, miss, or emphasize the distal scene (Fig. 1a). Synthetic labels also encode renderer choices for ray termination and alpha compositing, as in Hypersim [25]. When such sources are mixed in single-depth training, the resulting supervision collapses layered geometry into a dataset-shaped scalar target: as illustrated in Fig. 1b, either a domain-specific transparent-depth pipeline or a single-prediction foundation model must report one layer, even though the scene admits multiple valid depths. Thus, a model

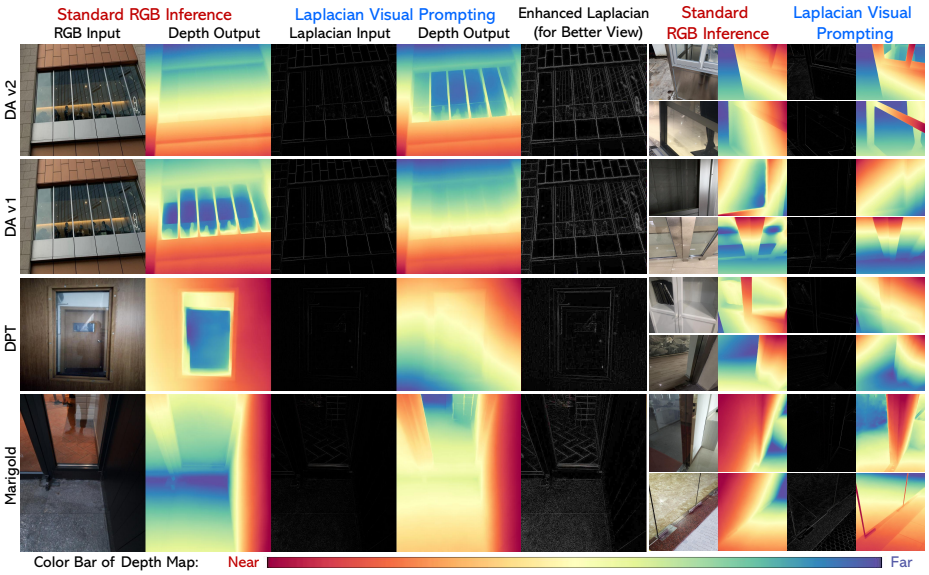


Fig. 2: Model-dependent depth-layer modulation. Standard RGB input reveals each model's default depth layer preference (**Cols. 2 and 7**). Laplacian Visual Prompting can change the reported layer for receptive frozen models [17, 23, 34, 35] and produce a candidate complementary depth hypothesis in ambiguous regions (**Cols. 4 and 9**).

does not learn layer-free geometry; it learns a **depth-layer preference**, the layer it tends to report when one ray is forced into one target.

The same convention affects evaluation. A standard single-layer metric rewards agreement with the recorded surface and can penalize another visible, physically valid layer. Thus, under layered ambiguity, the key question is not only *which prediction matches the label*, but which valid layer a model reports and how dataset bias shapes that behavior. To this end, we introduce `MultiDepth-3k` (`MD-3k`), a sparse ordinal benchmark for measuring depth-layer preference and multi-layer spatial relationship accuracy without dense metric ground truth. `MD-3k` annotates the transparent foreground and visible background with paired ordinal spatial relations, allowing us to evaluate *whether a biased single-layer depth output satisfies the foreground or the background spatial relation.*

Once the default model-specific depth-layer preference under RGB image inputs is measurable, we ask a second question: *can a frozen model's predicted layer be modulated by changing only the input representation?* This leads to Laplacian Visual Prompting (LVP), a simple and deterministic high-frequency input-space transform. Surprisingly, LVP can strongly change the predicted layer for certain frozen models, revealing candidate complementary ordinal behavior without retraining (Fig. 1d). Figure 2 qualitatively illustrates this modulation.

Using `MD-3k`, we first find that leading depth foundation models exhibit diverse layer preferences under standard RGB input. Beyond this default behavior, LVP reveals a surprising input-dependent modulation: it strongly changes the predicted

layer for DPT [23], Depth Pro [6], and several DAv2 models [35], while other models remain closer to their RGB preference. The strongest case is DAv2-L, whose RGB/LVP pair reaches 75.5% Multi-Layer Spatial Relationship Accuracy (ML-SRA) on `MD-3k`, above the strict 56.4% duplicated single-hypothesis ceiling. These observations make the central message concrete: even with fixed weights and a single-output head, a depth foundation model can express different valid ordinal relations when the input representation changes.

In summary, this framing leads to three contributions:

- We frame transparent-scene depth as a controlled case of **ambiguous layered 3D representation**, where multiple visible and geometrically valid depths may coexist along a ray, but a single-output depth model must choose one. We characterize this choice as a **model-intrinsic depth-layer preference**. To the best of our knowledge, we provide the first systematic characterization of this preference across diverse monocular depth foundation models.
- We introduce `MD-3k`, a real-world transparent-scene benchmark that makes depth-layer preference measurable. `MD-3k` provides sparse ordinal labels for two valid ray-wise depth layers, the transparent foreground surface and the visible background, enabling **per-layer and multi-layer evaluation**.
- We identify a surprising **input-dependent modulation** of depth-layer preference. Laplacian Visual Prompting serves as a training-free spectral probe that can expose complementary ordinal behavior in certain frozen models, while remaining model-dependent.

2 Related Work

Monocular depth estimation. Our work builds upon the generalization capabilities of modern Monocular Depth Estimation (MDE) foundation models. The field has evolved from domain-specific architectures trained on datasets such as KITTI [15] and NYUv2 [3, 12, 28] to general-purpose systems trained on large-scale mixed data. Current state-of-the-art models achieve robust zero-shot performance through diverse supervision strategies, including mixing heterogeneous datasets [5, 23, 24, 36], distilling generative priors from diffusion models [14, 16, 17, 26], or leveraging large-scale pseudo-labeling [34, 35]. Despite their robust representations, these models are architecturally constrained to produce a single depth value per pixel. This design choice requires models to resolve complex scene geometries into a single map, often resulting in systematic layer preferences under ambiguity. Our work investigates whether such layer preferences are influenced by input frequency content, and whether spectral input transformations can modulate which depth hypothesis a frozen model produces.

Depth estimation for ambiguous scenes. Recovering 3D scene geometry in the presence of transparent or specular surfaces is a persistent challenge. Early approaches formulated this as a completion problem, employing specialized layers to infer missing depth values [27, 38] or to regress background depth via transparency-aware losses [9, 13]. While effective in controlled settings, these methods do not explicitly represent multiple geometric interpretations within a single

model output. Recently, the field has moved toward multi-layer inference. Wen et al. [33] introduced a multi-layer synthetic dataset and fine-tuned depth models to predict separate geometric layers. These approaches occupy a fundamentally different research axis from ours: they ask how accurately a retrained model can predict multiple layers, whereas we ask how a frozen single-output model's expressed depth-layer preference changes under controlled input modulation. Our goal is diagnostic: we probe frequency-conditioned layer preferences in frozen models, rather than maximizing multi-layer prediction accuracy.

Visual prompting for model adaptation. We leverage Visual Prompting (VP) [1,2] to operate within the frozen model's input space. Inspired by prompting in NLP [7], VP has been adapted for vision-language alignment [18, 29, 31, 32], domain adaptation [10, 20], and adversarial robustness [8]. Most existing VP methods focus on learnable prompts such as optimized pixel patches or input tokens. A distinct sub-category is input-space prompting [30], where the input signal is transformed to steer model behavior without any parameter optimization. Our work extends this paradigm to 3D geometry. We apply non-learnable spectral input-space prompting to analyze and modulate depth-layer preference in foundation models, demonstrating that simple spectral prompting can alter the depth hypothesis produced by a frozen single-output estimator.

3 Methodology

We define a framework for measuring depth-layer preference and testing whether this preference can be modulated in frozen single-output depth models. The framework has three parts. First, we define the single-output setting and the two-layer ordinal representation used for ambiguous scenes. Second, we formulate depth-layer preference for a single prediction and paired-hypothesis complementarity for two predictions. Third, we apply Laplacian Visual Prompting (LVP) as a training-free input transformation and compare its outputs with standard RGB outputs under the same ordinal evaluation.

3.1 Preliminaries: From Single-Layer Depth to Layered Geometry

The single-layer constraint. We consider a pre-trained monocular depth model $f_\theta : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W}$, parameterized by frozen model weights θ, that produces a single depth estimate $\hat{\mathcal{D}} = f_\theta(\mathcal{I})$ per input image $\mathcal{I}$. Regardless of architecture, all models evaluated in this work emit a scalar depth value per pixel at inference time. For opaque scenes this representation is usually well posed. In ambiguous layered scenes, however, one ray can contain evidence for multiple visible surfaces. We use transparency as a clean two-layer case and denote the valid interpretations as $\mathcal{D}^{(1)}$ for the transparent foreground and $\mathcal{D}^{(2)}$ for the visible background. A single scalar output cannot express both layers, so it must report one layer convention rather than an intrinsically unique depth.

Multi-layer geometry via sparse ordering. We represent the two-layer scene by an ordered pair $(\mathcal{D}^{(1)}, \mathcal{D}^{(2)})$ such that $\forall\, \mathbf{x},\ \mathcal{D}^{(1)}(\mathbf{x}) \le \mathcal{D}^{(2)}(\mathbf{x})$, where $\mathbf{x}$

denotes a spatial location on the image plane. Reliable dense multi-layer depth is fundamentally difficult to collect in real layered scenes. Physical sensors generally return one physical layer, or fail, penetrate, or reflect in material-dependent ways, rather than capturing the complete stack of visible surfaces along a ray. A dense metric label is therefore not sensor-neutral for transparency. Following DIW [11] and `DA-2K` [35], we formulate correctness through sparse ordinal constraints.

Let $\mathcal{P} = \{(\mathbf{u}_m, \mathbf{v}_m)\}_{m=1}^{M}$ be a set of sparse point pairs sampled from ambiguous regions. We sample one pair per image and manually annotate its ordinal relation for both the foreground and background layers. The spatial order of ground truth for layer $k \in \{1, 2\}$ is $y_m^{(k)} = \operatorname{sign}\big(\mathcal{D}^{(k)}(\mathbf{u}_m) - \mathcal{D}^{(k)}(\mathbf{v}_m)\big)$.

A predicted depth map $\hat{\mathcal{D}}$ is considered valid for layer k if its relative depth ordering matches the corresponding ordinal label. For clarity, we write $\hat{\mathcal{D}} \equiv y_m^{(k)}$ as shorthand for $\operatorname{sign}\big(\hat{\mathcal{D}}(\mathbf{u}_m) - \hat{\mathcal{D}}(\mathbf{v}_m)\big) = y_m^{(k)}$ in the following paragraphs.

3.2 Problem Formulation: Layer Preference and Paired Hypotheses

Single-output depth-layer preference. The sparse ordinal formulation lets us measure the layer convention expressed by a single-output model. Depth-layer preference asks which valid layer a model tends to report under ambiguity. We define $\alpha(f_\theta)$ as the expected difference in sparse ordinal correctness between the background and foreground layers for a frozen model:

$$\alpha(f_\theta) = \mathbb{E}_{m\sim\mathcal{P}}\Big[\mathbb{I}(\hat{\mathcal{D}} \equiv y_m^{(2)}) - \mathbb{I}(\hat{\mathcal{D}} \equiv y_m^{(1)})\Big], \tag{1}$$

where $\alpha > 0$ indicates background preference and $\alpha < 0$ indicates foreground preference. The absolute value of α denotes preference strength.

Paired-hypothesis complementarity. A single depth map can report only one layer convention. To evaluate whether two depth outputs provide complementary evidence, we form a candidate depth pair $\mathbf{H} = \{\hat{\mathcal{D}}_A, \hat{\mathcal{D}}_B\}$ and jointly evaluate its two maps against both annotated depth layers. We define success as the existence of a dataset-level permutation $\pi^\star$ that maps hypotheses to layers and maximizes the joint satisfaction of the sparse constraints:

$$\pi^\star = \arg\max_{\pi\in S_2} \sum_{m=1}^{M} \mathbb{I}\left(\bigwedge_{k=1}^{2} \hat{\mathcal{D}}_{\pi(k)} \equiv y_m^{(k)}\right). \tag{2}$$

Benchmark labels calibrate a single dataset-level assignment for each candidate pair. The assignment is fixed across all images, so paired evaluation measures candidate-pair complementarity after dataset-level label matching.

3.3 Scope of Ambiguity

Radiometric superposition vs. occlusion. We focus on *transparency* as a primary mode of ambiguity. Unlike **occlusion**, where the background signal is physically blocked, transparency results in **radiometric superposition**: photons

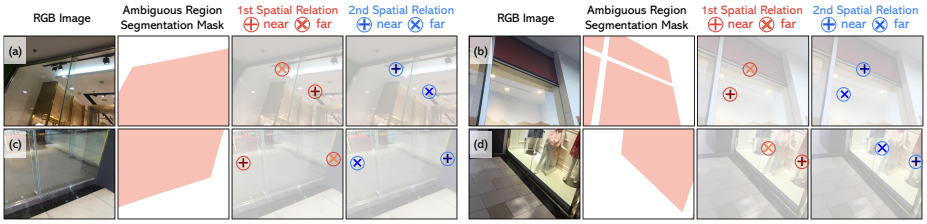


Fig. 3: `MD-3k` benchmark. Examples showing ambiguous region masks and sparse point pairs with multi-layer spatial labels. Spatial relationships of sparse point pairs reverse across layers in (**a**)–(**c**) and remain consistent in (**d**).

from both the foreground surface and the background contribute to the observed pixel intensity. Consequently, visual cues from multiple surfaces coexist in the input signal $\mathcal{I}$ in superimposed form. Our method does not attempt to hallucinate missing geometry. Instead, it probes how frozen models respond when frequency components of this signal are selectively emphasized in the image.

3.4 `MultiDepth-3k`: Sparse Ordinal Benchmark for Ambiguity

Building on the sparse ordinal formulation, `MultiDepth-3k` (`MD-3k`) provides a real-world benchmark for measuring layer choice in transparent scenes. It consists of 3,161 RGB images sourced from the GDD dataset [19], with one annotated point pair per image. Because physical sensing cannot reliably recover a dense, sensor-neutral stack of visible layers, we use sparse ordinal relations rather than dense metric multi-layer depth, following DIW [11] and `DA-2K` [35], as shown in Fig. 3. Each pair has labels for both layers. Masks and labels were cross-checked by multiple annotators in multiple review rounds before evaluation.

Benchmark statistics. As shown in Fig. 4, the dataset captures diverse ambiguity ratios. We partition the dataset into two subsets. In the ***Same*** subset (1,783 pairs), relative depth ordering is consistent across layers. In the ***Reverse*** subset (1,378 pairs), the two layers impose conflicting orders. Consequently, a single-output model cannot satisfy both layers on the *Reverse* subset by duplicating one map, which is exactly the case where multiple hypotheses become necessary. Because all images are drawn from GDD [19], broader validation across capture domains, transparent materials, and object categories remains necessary before treating the benchmark as general transparent-scene coverage rather than a focused diagnostic for measuring layer choice under transparency.

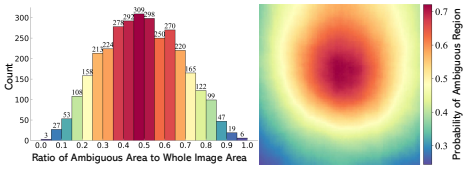


Fig. 4: Statistics. Left: distribution of ambiguous area ratio per image. **Right**: 2D spatial heatmap of ambiguous regions over benchmark images in `MD-3k`, shown in normalized image coordinates.

3.5 Evaluation Metrics

Given these two-layer ordinal labels, we report three quantities. Together, they separate single-output layer preference from paired-output complementarity.

Spatial Relationship Accuracy (SRA). For a given layer $i \in \{1, 2\}$, SRA(i) measures the fraction of point pairs for which the predicted ordering is consistent with the ground truth: $\mathrm{SRA}(i) = \frac{1}{|\mathcal{P}|} \sum_{m=1}^{M} \mathbb{I}\left(\hat{\mathcal{D}} \equiv y_m^{(i)}\right)$.

Depth-Layer Preference (α). We diagnose the direction of a model's bias using the empirical estimator of Eq. (1): $\alpha(f_\theta) = \mathrm{SRA}(2) - \mathrm{SRA}(1)$. A positive α indicates background preference. A negative α indicates foreground preference.

Multi-Layer Spatial Relationship Accuracy (ML-SRA). ML-SRA measures the fraction of point pairs for which the relative depth ordering is correctly predicted for both depth layers: $\text{ML-SRA} = \frac{1}{|\mathcal{P}|} \sum_{m=1}^{M} \mathbb{I}\left(\bigwedge_{k=1}^{2} \hat{\mathcal{D}}_{\pi^\star(k)} \equiv y_m^{(k)}\right)$. The benchmark-level assignment $\pi^\star$ from Eq. (2) is fixed for every image. In our RGB/LVP evaluation (Sec. 4.2), we use a deterministic calibration rule: the RGB output is assigned to the layer for which its benchmark-level SRA is higher, and the LVP output is assigned to the complementary layer.

3.6 Laplacian Visual Prompting: Querying the Expressed Convention

After defining how layer choice is measured, we use LVP as the complementary query: can the same frozen model express a different convention under a different input representation? As illustrated in Fig. 5a–b, standard depth foundation model training couples RGB images to single-layer supervision, so RGB inference returns one preferred/biased depth layer under ambiguity. LVP constructs a Laplacian-transformed RGB image from the same image and uses it as an alternative input representation. As shown in Fig. 5c, this can alter the expressed depth-layer preference in architecturally sensitive models without modifying frozen model weights θ.

Spectral reweighting. We derive the discrete operator from the continuous Laplacian $\Delta\mathcal{I} = \nabla^2\mathcal{I}$. Using a second-order finite difference approximation, we obtain the discrete Laplacian kernel $\mathcal{K}_\mathcal{L}$:

$$\mathcal{K}_\mathcal{L} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{3}$$

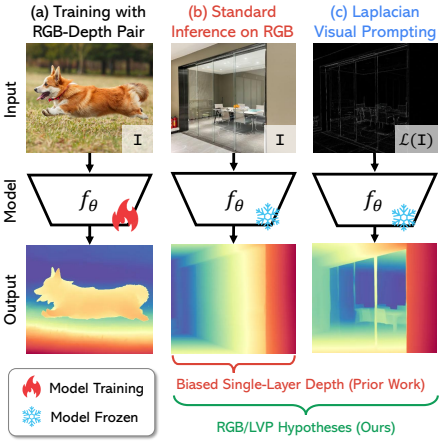


Fig. 5: The Laplacian Visual Prompting (LVP) method. (**a**) Standard model training couples RGB to single-layer depth. (**b**) At inference, the standard RGB input yields a single depth estimate, which is biased for ambiguous scenes. (**c**) LVP transforms the input via per-channel floating-point convolution with the Laplacian kernel, followed by min–max mapping back to the image-input value range, producing a candidate alternative output hypothesis from the same frozen depth model.

We apply this convolution channel-wise in floating point to the input RGB image $\mathcal{I}$, yielding signed residuals $\mathcal{R}_{\text{raw},c} = \mathcal{I}_c * \mathcal{K}_{\mathcal{L}}$ for each color channel c.

Normalization and inference. The raw Laplacian response is a signed floating-point residual and is not directly a valid image input. We therefore map it back to the value range of the model's image-input representation before applying the model-specific preprocessing pipeline. Let this image-input range be $[a, b]$. For example, $[a, b] = [0, 1]$ for floating-point RGB images, or equivalently $[a, b] = [0, 255]$ before conversion by an image processor. We define

$$\mathcal{L}(\mathcal{I})_c(\mathbf{x}) = a + (b-a)\frac{\mathcal{R}_{\text{raw},c}(\mathbf{x}) - \min_{\mathbf{x}'} \mathcal{R}_{\text{raw},c}(\mathbf{x}')}{\max_{\mathbf{x}'} \mathcal{R}_{\text{raw},c}(\mathbf{x}') - \min_{\mathbf{x}'} \mathcal{R}_{\text{raw},c}(\mathbf{x}') + \epsilon}, \tag{4}$$

where the minimum and maximum are computed spatially for each channel c, and ϵ is a small constant (e.g., $\epsilon = 10^{-8}$) for numerical stability. This step is an image-space value-range mapping. It does not replace or modify the model-specific resizing, rescaling, or normalization used for standard RGB inference. In our implementation, the image processor expects a standard 8-bit RGB image, so we map the LVP residual image back to the `uint8` value range $[0, 255]$ and pass the resulting image through the same processor used for the original RGB input.

Given a frozen model f_θ, we obtain the LVP-conditioned depth hypothesis:

$$\mathcal{D}_{\text{LVP}} = f_\theta(\mathcal{L}(\mathcal{I}))\,. \tag{5}$$

For ML-SRA, $(\mathcal{D}_{\text{RGB}}, \mathcal{D}_{\text{LVP}})$ is treated as an unordered candidate pair. Benchmark labels select one global permutation $\pi^\star$ for each model's output pair, and that assignment is fixed for all images. No per-instance oracle is used. Thus, ML-SRA measures pair complementarity after dataset-level label matching, not automatic layer control. Deployment would require labeled calibration or an external semantic or uncertainty-based selector.

4 Experiments

Research question. We first use `MD-3k` to measure the default depth-layer preference of leading models under standard RGB image input. We then ask the question: can a training-free input transformation modulate that preference in a frozen model? We evaluate this through RGB/LVP candidate pairs, report ML-SRA against the two layer-specific ordinal relations, and analyze when the modulation appears or fails. Accordingly, we report (i) depth-layer preference, (ii) ML-SRA for RGB/LVP candidate pairs, (iii) scale and training-distribution effects, (iv) prompt-design ablation studies, and (v) uses of paired hypotheses.

Experimental setup. We evaluate a diverse suite of pre-trained monocular depth models in a strictly training-free manner. The benchmarked models include the Depth Anything series (DAv1/v2-{S,B,L}) along with the domain-specialized Indoor/Outdoor variants of DAv2 (DAv2-I/O-{S,B,L}) [34, 35]. We also evaluate discriminative architectures (DPT [23], ZoeDepth [4]), generative models (Marigold [17], GeoWizard [14]), and metric estimators (Depth Pro [6], UniK3D [21], UniDepth-v2 [22]). Our evaluations are mainly performed on `MD-3k`.

Table 1: Per-layer Spatial Relationship Accuracy (SRA) [%] on `MD-3k` and reference SRA on `DA-2K`. On `MD-3k`, SRA(1) and SRA(2) measure agreement between a model's single-layer depth prediction and the transparent foreground layer or visible background layer, respectively. The *Reverse* subset contains conflicting foreground/background ordinal relations, while the *Same* subset contains consistent relations and is therefore reported as SRA(1/2). `DA-2K` is included as a non-ambiguous reference. Bold entries mark the better RGB or LVP input for the corresponding foreground/background SRA comparison on the *Overall* and *Reverse* subsets.

Method	(a) `MD-3k` (*Overall*)				(b) `MD-3k` (*Reverse*)				(c) `MD-3k` (*Same*)		(d) `DA-2K`	
	RGB Input		LVP Input		RGB Input		LVP Input		RGB	LVP	RGB	LVP
	SRA(1)	SRA(2)	SRA(1)	SRA(2)	SRA(1)	SRA(2)	SRA(1)	SRA(2)	SRA(1/2)	SRA(1/2)	SRA	SRA
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Marigold	70.7	**79.9**	**70.8**	77.4	39.6	**60.4**	**42.3**	57.7	95.0	92.6	88.9	78.9
GeoWizard	68.3	**83.8**	**70.3**	78.7	32.3	**67.7**	**40.3**	59.7	96.2	93.4	90.3	85.0
ZoeDepth	58.7	**92.6**	**74.6**	70.9	12.0	**88.0**	**54.3**	45.7	94.9	90.4	86.7	77.2
DPT	58.0	**91.9**	**75.7**	71.1	10.2	**89.8**	**55.2**	44.8	94.8	91.5	83.2	72.2
Depth Pro	**84.9**	69.6	73.5	**76.8**	**67.6**	32.4	46.2	**53.8**	98.3	94.6	95.8	84.1
UniDepth-v2-L	**66.8**	**86.6**	65.5	85.4	**27.3**	72.7	27.1	**72.9**	97.4	95.1	95.4	92.2
UniK3D-L	64.3	**88.9**	**67.1**	84.3	21.8	**78.2**	**30.3**	69.7	97.2	95.6	93.1	85.8
DAv1-S	56.8	**95.3**	**60.6**	85.4	5.8	**94.2**	**21.5**	78.5	96.1	90.7	88.7	83.0
DAv1-B	56.8	**95.4**	**58.8**	89.7	5.8	**94.2**	**14.6**	85.4	96.3	93.0	89.9	86.2
DAv1-L	56.6	**96.0**	**59.2**	90.9	4.8	**95.2**	**13.6**	86.4	96.6	94.3	89.5	88.5
DAv2-O-S	71.6	**77.3**	**81.4**	60.1	43.4	**56.6**	**74.4**	25.6	93.3	86.8	82.3	60.9
DAv2-O-B	70.3	**80.3**	**77.2**	65.8	38.5	**61.5**	**63.1**	36.9	94.8	88.2	89.7	70.4
DAv2-O-L	70.4	**82.4**	**71.5**	79.5	36.2	**63.8**	**40.8**	59.2	96.7	95.2	93.7	83.8
DAv2-I-S	**80.4**	71.2	70.0	**74.3**	**60.6**	39.4	45.1	**54.9**	95.8	89.3	88.5	79.2
DAv2-I-B	**83.1**	69.7	72.6	**76.1**	**65.3**	34.7	46.0	**54.0**	96.8	93.1	91.8	81.1
DAv2-I-L	**85.2**	68.1	68.1	**82.6**	**69.6**	30.4	33.5	**66.5**	97.3	95.0	94.8	88.4
DAv2-S	**78.0**	76.2	61.5	**85.8**	**52.0**	48.0	22.1	**77.9**	98.0	91.9	95.1	86.6
DAv2-B	**82.4**	72.3	60.5	**88.6**	**61.7**	38.3	17.7	**82.3**	98.5	93.5	96.7	89.5
DAv2-L	**84.0**	70.6	60.2	**89.9**	**65.3**	34.7	15.9	**84.1**	98.5	94.4	96.9	91.5

4.1 Depth-Layer Preference Analysis

Per-layer ordinal behavior. Before aggregating RGB/LVP outputs into paired depth hypotheses, we first report the single-output depth layer behavior under RGB and LVP inputs directly. Table 1 presents per-layer Spatial Relationship Accuracy (SRA) with respect to the transparent foreground layer and the visible background layer on `MD-3k`, together with `DA-2K` results as a non-ambiguous reference. On the *Same* subset of `MD-3k`, where the two layers induce consistent ordinal spatial relations, SRA is generally on par with `DA-2K`, a standard benchmark of mostly non-ambiguous scenes. This confirms that `MD-3k` does not merely introduce a harder ordinal task; its distinctive challenge lies in the *Reverse* subset, where the two valid layers impose conflicting spatial relations.

Heterogeneity of RGB bias. Table 1 reports the raw per-layer ordinal agreement, from which the depth-layer preference is computed as $\alpha = \mathrm{SRA}(2) - \mathrm{SRA}(1)$. The RGB circles in Fig. 6 visualize this derived preference on the *Reverse* subset. Depth foundation models exhibit strong but inconsistent layer preferences under standard RGB image input. The *Depth Anything* family shows this split: general-

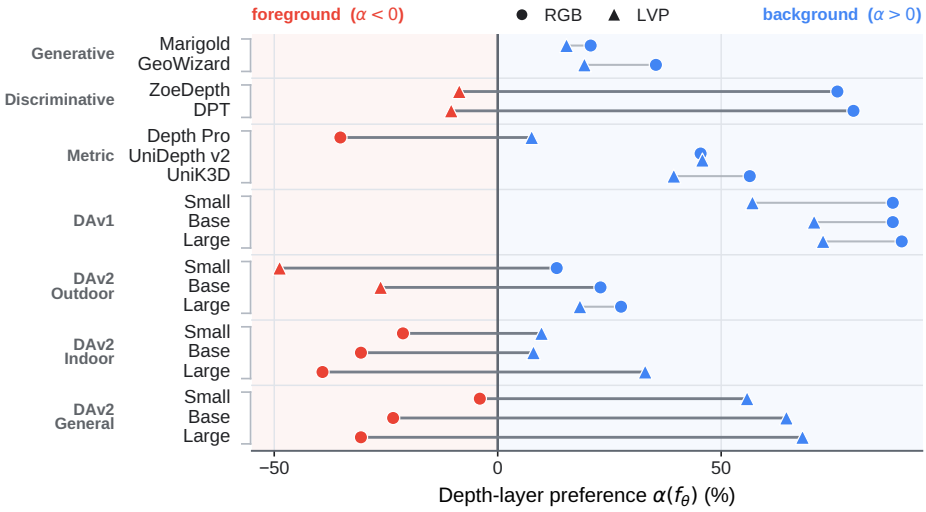


Fig. 6: Model-dependent depth-layer preference. On `MD-3k` *Reverse*, each row links RGB (circle) and LVP (triangle). Fill indicates the preferred layer (red: foreground; blue: background), and crossing $\alpha = 0$ indicates a layer change. The varied endpoints and shifts expose model-specific RGB priors and LVP responses.

purpose DAv2 (DAv2-S/B/L) and indoor-tuned DAv2-I variants favor the first layer (transparent foreground, $\alpha < 0$), whereas outdoor-tuned DAv2-O variants and DAv1 favor the second layer (background, $\alpha > 0$), similar to generative models such as Marigold. This pattern matches the supervision story above: mixed training data and domain bias affect which surface a model reports under transparency. In particular, indoor-tuned variants more often select the proximal surface, whereas outdoor-tuned variants more often select the distal scene.

Surprising LVP modulation. The same per-layer SRA values in Table 1 also allow us to compute how LVP changes the derived preference $\alpha = \mathrm{SRA}(2) - \mathrm{SRA}(1)$. The LVP triangles and RGB-to-LVP segments in Fig. 6 visualize this preference shift on the *Reverse* subset. The effect is strongly model-specific and, for several models, unexpectedly large: general-purpose DAv2, DPT, ZoeDepth, and Depth Pro shift substantially, whereas DAv1 and several generative estimators respond weakly. For example, on the *Reverse* subset, DAv2-L changes from foreground-biased RGB behavior to background-biased LVP behavior: SRA(1) drops from 65.3% to 15.9%, while SRA(2) rises from 34.7% to 84.1%. This contrast suggests that LVP is not a universal layer switch, but a probe of spectral receptivity: it reveals which frozen backbones can express alternative layer behavior under the same input-level intervention.

4.2 Multi-Layer Ordinal Performance

Quantitative analysis. Having diagnosed single-output preference, we next evaluate paired-output complementarity. Table 2 reports ML-SRA for the RGB/LVP

Table 2: ML-SRA [%] on `MD-3k`. † The strict 56.4% ceiling applies to a duplicated single-map pair under ML-SRA and follows from the benchmark partition ratio.

Method	*Overall*	*Reverse*	*Same*
Ideal Collapsed Baseline†	56.4†	0.0	100.0†
Marigold	57.4	15.3	89.8
GeoWizard	59.5	17.6	91.9
ZoeDepth	68.8	45.4	86.8
DPT	70.2	46.4	88.7
Depth Pro	66.3	31.1	93.5
UniDepth-v2-L	61.3	13.6	93.7
UniK3D-L	58.9	19.2	**93.9**
DAv1-S	57.9	17.7	89.0
DAv1-B	56.6	11.4	91.5
DAv1-L	57.1	10.9	92.8
DAv2-O-S	63.0	36.6	83.5
DAv2-O-B	62.7	32.9	85.6
DAv2-O-L	60.4	17.6	93.4
DAv2-I-S	60.9	27.7	86.5
DAv2-I-B	63.7	28.1	91.2
DAv2-I-L	71.1	42.5	93.2
DAv2-S	67.2	36.9	90.7
DAv2-B	73.3	48.2	92.7
DAv2-L	**75.5**	**52.2**	93.6

Table 3: Contextual comparison with semantic priors. LVP uses DAv2-L; mask interpolation is built on background-biased DAv1-L to obtain a distal map before estimating transparent regions. ‡ Using a transparency mask predictor [19] with mIoU 0.88 on `MD-3k`.

Method	Sem.	*Overall*	*Reverse*	*Same*
LVP (DAv2-L)	No	75.5	52.2	93.6
Mask interpolation (pred., DAv1-L)‡	Yes	75.8	55.2	91.6
Mask interpolation (GT, DAv1-L; oracle)	Yes	82.5	69.2	92.7

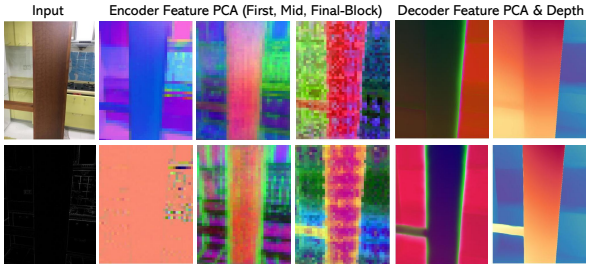
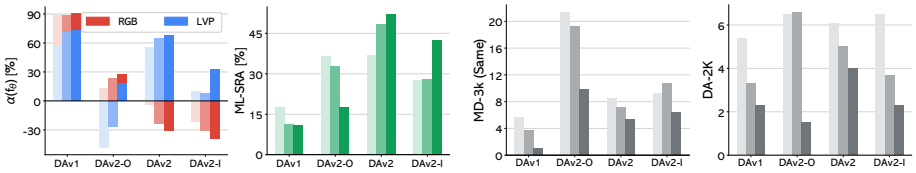


Fig. 7: Feature visualization. PCA of DAv2-L encoder and decoder features. Under LVP input (Bottom), activations place greater emphasis on background high-frequency edges than under RGB input (Top). This qualitative, input-dependent feature highlighting is not evidence of discrete latent depth layers.

candidate pair across all models. To contextualize these results, we define an *Ideal Collapsed Baseline*: a hypothetical model that perfectly predicts the primary RGB depth but naïvely duplicates it for the secondary layer. This baseline achieves 100% on the *Same* subset but **0%** on the *Reverse* subset because satisfying conflicting ordinal constraints with a single depth map is impossible by construction. Its weighted overall score of 56.4% (derived from the benchmark partition: $1783/(1783 + 1378)$) therefore represents the strict ceiling for any duplicated single-map pair under ML-SRA. Our method with DAv2-L achieves **75.5%** overall and **52.2%** on the *Reverse* subset, representing a jump from 0% to 52.2% in precisely the cases where a duplicated single output cannot satisfy both relations. Therefore, the **+19.1-point** gain over the ideal single-hypothesis ceiling establishes the central empirical claim: the same frozen model, queried by RGB and LVP inputs, can jointly satisfy contradictory ordinal constraints beyond what any single depth map can achieve.

Model-family trend. Table 2 also shows that LVP response is model-dependent, but not explained by a simple discriminative/generative split. DAv2 and DPT exhibit strong modulation, whereas DAv1 and diffusion-based estimators such as Marigold and GeoWizard respond weakly. This suggests that LVP effectiveness



(a) *Reverse* subset of `MD-3k` (b) *Same* subset of `MD-3k` and `DA-2K`

Fig. 8: Scaling analysis. (**a**) On the *Reverse* subset of `MD-3k`, larger variants benefit most when RGB and LVP select different layers; when both inputs share the same layer bias, the candidate pair offers less complementarity. (**b**) On *Same* subset of `MD-3k` and `DA-2K`, we plot the RGB/LVP ML-SRA gap in percentage points. Smaller bars mean that LVP stays closer to the RGB baseline when the ordinal relation is consistent.

depends on the model's spectral receptivity, shaped by both architecture and training regime. We leave a direct causal analysis of this behavior to future work.

Comparison with semantic priors. As a complementary reference, Table 3 compares against a semantics-assisted pipeline. LVP uses DAv2-L, whereas mask-assisted interpolation uses DAv1-L because its background bias supplies a distal map from which the transparent region is interpolated. The pipelines therefore differ in both backbone and prior. The predicted-mask pipeline reaches 75.8%, and its GT-mask oracle reaches 82.5%, but both require semantic localization and planar interpolation. In contrast, LVP reaches 75.5% without an auxiliary semantic segmentation model on its stated backbone.

Latent feature visualization. Finally, Figure 7 provides a qualitative view of input-dependent feature highlighting: under LVP, activations emphasize high-frequency structures that differ from those emphasized by RGB. This feature-level change supports the spectral-modulation interpretation, while the quantitative claim of the paper rests on the ordinal behavior measured by `MD-3k`.

4.3 Ablation and Prompt Analysis

Scaling behavior. Figure 8 further separates complementarity from stability. On the *Reverse* subset (Fig. 8a), scaling helps when RGB and LVP move toward different layers, as in DAv2. In contrast, when the two inputs retain the same layer bias (most clearly in DAv2-O-L), the paired output has less room to satisfy conflicting relations. In the consistent-relation setting (Fig. 8b), the y-axis is the RGB/LVP ML-SRA gap. As model size grows, this gap often shrinks on the *Same* subset of `MD-3k` and `DA-2K`, indicating that LVP perturbs the RGB baseline less when the ordinal relation across layers is stable.

Prompt design ablation. To isolate the role of the input transform, we compare LVP against a low-pass Gaussian prompt in Table 4. On the *Reverse* subset, the Gaussian prompt fails across all models (near-zero scores), as low-frequency preservation merely retains the primary depth hypothesis. LVP succeeds, supporting high-frequency emphasis as an operational factor associated with preference modulation. This trade-off is spectrally asymmetric: on the *Same* subset, the

Table 4: Laplacian (LVP) vs. Gaussian (GAU) prompts via ML-SRA [%].

Input	Marigold	GeoWizard	ZoeDepth	DPT	DAv1-S	DAv1-B	DAv1-L	DAv2-O-S	DAv2-O-B	DAv2-O-L	DAv2-I-S	DAv2-I-B	DAv2-I-L	DAv2-S	DAv2-B	DAv2-L
(a) ***Reverse*** **Subset of** `MD-3k`																
LVP	**15.3**	**17.6**	**45.4**	**46.4**	**17.7**	**11.4**	**10.9**	**36.6**	**32.9**	**17.6**	**27.7**	**28.1**	**42.5**	**36.9**	**48.2**	**52.2**
GAU	6.2	6.0	0.7	0.3	0.7	0.3	0.4	5.7	2.6	0.6	1.0	0.7	2.4	3.6	4.6	4.5
(b) ***Same*** **Subset of** `MD-3k`																
LVP	89.8	91.9	86.8	88.7	89.0	91.5	92.8	83.5	85.6	93.4	86.5	91.2	93.2	90.7	92.7	93.6
GAU	**94.2**	**95.3**	**94.6**	**94.3**	**95.7**	**96.1**	**96.5**	**92.6**	**94.6**	**96.4**	**95.3**	**96.6**	**96.9**	**97.8**	**98.1**	**98.2**

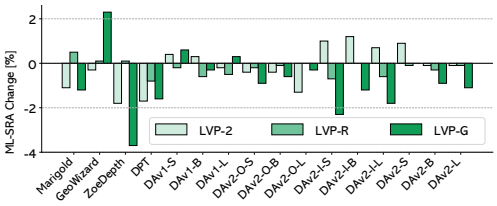


Fig. 9: Ablation of LVP design. Relative change in ML-SRA [%] compared to default LVP. Performance is robust to kernel variants (LVP-2: 8-neighbor), sign flip of Laplacian kernel (LVP-R), and grayscale input (LVP-G).

Table 5: High-frequency prompt comparison. ML-SRA [%] values are grouped in *Overall/Reverse/Same*.

Model	LVP	Sobel	Fourier	Wavelet
DAv2-S	67.2	69.8	66.9	68.7
	36.9	40.3	33.2	40.3
	90.7	92.6	93.0	90.6
DAv2-B	73.3	72.2	72.4	73.1
	48.2	44.5	43.0	47.8
	92.7	93.7	**95.2**	92.6
DAv2-L	**75.5**	73.9	75.0	74.8
	52.2	47.2	49.1	50.4
	93.6	94.5	95.0	93.7

Gaussian outperforms LVP across all models, since low-frequency preservation suffices for unambiguous scenes while high-frequency perturbation unnecessarily disrupts the primary hypothesis. This contrast separates two regimes: high-frequency prompting helps when the two layers impose conflicting orderings, whereas low-frequency preservation is better when one ordering is already sufficient. Performance is stable across kernel variations (4 vs. 8 neighbors), sign flips, and grayscale conversion (Fig. 9), which is consistent with sensitivity to broad high-frequency emphasis rather than dependence on one Laplacian discretization.

Alternative spectral prompts. We then test whether the effect is specific to the Laplacian operator. Table 5 evaluates Sobel, Fourier high-pass, and wavelet prompts. No operator dominates every model and subset: Sobel matches the best Reverse score for DAv2-S, while LVP is strongest for DAv2-B/L. The shared effect across continuous high-frequency operators supports the broader spectral-modulation finding; LVP remains a simple, parameter-free default rather than a uniquely privileged transform under this benchmark.

4.4 Downstream Applications

Conditional generation and video hypotheses. In Fig. 10, the RGB/LVP pair provides candidate depth conditions from the same frozen model rather than a final layered reconstruction. In conditional generation, selected hypotheses can drive ControlNet [37] renderings that emphasize different visible geometry while keeping the RGB scene fixed. In video, applying RGB and LVP frame by frame produces distinct depth streams, making layered ambiguity visible over time.

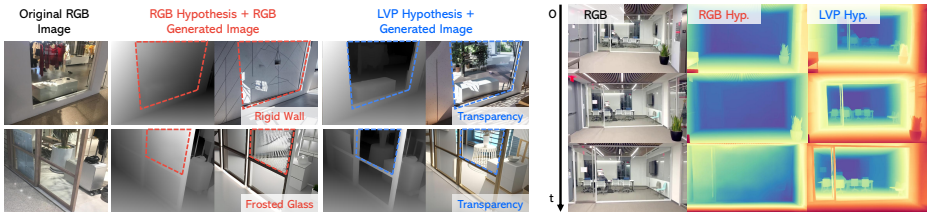


Fig. 10: Downstream illustrations. Selected RGB/LVP-conditioned depth hypotheses provide alternative ControlNet conditions and frame-wise depth streams.

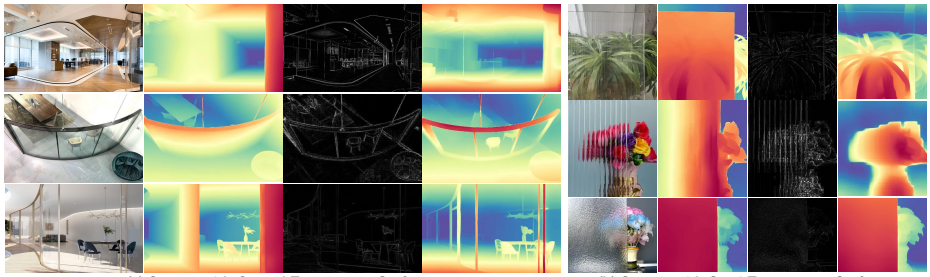


Fig. 11: Generalization and failure cases of LVP. LVP can modulate depth-layer preference in challenging curved-glass scenes, but may fail on semi-transparent surfaces where foreground and background cues are frequency-entangled.

5 Discussion and Conclusions

Ambiguous layered scenes reveal a fundamental limitation of single-depth estimation: a single scalar target collapses multiple geometrically valid ray-wise interpretations into one dataset-dependent layer convention. `MD-3k` makes this convention explicit. Leading depth foundation models exhibit diverse RGB layer preferences, while LVP shows that some frozen models can be modulated to express complementary layer hypotheses without retraining. These findings suggest that standard RGB inference captures only one slice of a richer geometric posterior, and that future depth systems should represent, evaluate, and learn from multiple plausible scene depths rather than treating ambiguity as noise.

Limitations and future work. As shown in Fig. 11, curved glass can preserve separable cues and produce a useful foreground shift, while textured transparent surfaces can entangle layer frequencies and fail. LVP is therefore model-dependent and should not be treated as a reliable layer extractor. `MD-3k` is sparsely-annotated and focused on transparent scenes, leaving dense validation, automatic layer selection, and broader ambiguous-scene benchmarks to future work.

Acknowledgments

The authors gratefully acknowledge Modal Labs for providing partial support through a generous academic compute grant.

References

1. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022)
2. Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A.L., Darrell, T., Malik, J., Efros, A.A.: Sequential modeling enables scalable learning for large vision models. In: CVPR. pp. 22861–22872 (2024)
3. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: CVPR (2021)
4. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv:2302.12288 (2023)
5. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv:2307.14460 (2023)
6. Bochkovskiy, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: ICLR (2025)
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS **33**, 1877–1901 (2020)
8. Chen, A., Lorenz, P., Yao, Y., Chen, P.Y., Liu, S.: Visual prompting for adversarial robustness. In: ICASSP. pp. 1–5. IEEE (2023)
9. Chen, K., Wang, S., Xia, B., Li, D., Kan, Z., Li, B.: TODE-Trans: Transparent object depth estimation with transformer. In: ICRA. pp. 4880–4886 (2023)
10. Chen, L., Fan, Y., Ye, Y.: Adversarial reprogramming of pretrained neural networks for fraud detection. In: CIKM. pp. 2935–2939 (2021)
11. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NeurIPS (2016)
12. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NeurIPS (2014)
13. Fang, H., Fang, H.S., Xu, S., Lu, C.: Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. IEEE Robotics and Automation Letters **7**(3), 7383–7390 (2022)
14. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. ECCV (2024)
15. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR (2013)
16. Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: DepthFM: Fast monocular depth estimation with flow matching. In: AAAI (2025)
17. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024)
18. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. CVPR (2023)
19. Mei, H., Yang, X., Wang, Y., Liu, Y., He, S., Zhang, Q., Wei, X., Lau, R.W.: Don't hit me! glass detection in real-world scenes. In: CVPR. pp. 3687–3696 (2020)
20. Neekhara, P., Hussain, S., Du, J., Dubnov, S., Koushanfar, F., McAuley, J.: Cross-modal adversarial reprogramming. In: WACV. pp. 2427–2435 (2022)

21. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Van Gool, L.: Unik3d: Universal camera monocular 3d estimation. In: CVPR. pp. 1028–1039 (2025)
22. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. arXiv preprint arXiv:2502.20110 (2025)
23. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
24. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. TPAMI (2022)
25. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021)
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
27. Sajjan, S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., Song, S.: ClearGrasp: 3D shape estimation of transparent objects for manipulation. In: ICRA. pp. 3634–3642 (2020)
28. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
29. Singha, M., Pal, H., Jha, A., Banerjee, B.: AD-CLIP: Adapting domains in prompt space using CLIP. In: ICCV Workshops. pp. 4355–4364 (2023)
30. Tsai, Y.Y., Chen, P.Y., Ho, T.Y.: Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. ICML (2020)
31. Wang, H., Liu, F., Jiao, L., Wang, J., Hao, Z., Li, S., Li, L., Chen, P., Liu, X.: Vilt-clip: Video and language tuning clip with multimodal prompt learning and scenario-guided optimization. In: AAAI (2024)
32. Wasim, S.T., Naseer, M., Khan, S., Khan, F.S., Shah, M.: Vita-clip: Video and text adaptive clip via multimodal prompting. In: CVPR. pp. 23034–23044 (2023)
33. Wen, H., Zuo, Y., Subramanian, V., Chen, P., Deng, J.: Seeing and seeing through the glass: Real and synthetic data for multi-layer depth estimation. ICCV (2025)
34. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
35. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. NeurIPS (2024)
36. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: ICCV (2023)
37. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
38. Zhu, L., Mousavian, A., Xiang, Y., Mazhar, H., van Eenbergen, J., Debnath, S., Fox, D.: RGB-D local implicit function for depth completion of transparent objects. In: CVPR. pp. 12725–12734 (2021)