MULTI-HYPOTHESIS SPATIAL FOUNDATION MODEL: RETHINKING AND DECOUPLING DEPTH AMBIGUITY VIA LAPLACIAN VISUAL PROMPTING

Xiaohao Xu[1] **Feng Xue**[1] **Xiang Li**[2] **Haowei Li**[1] **Shusheng Yang**[3]
Tianyi Zhang[2] **Matthew Johnson-Roberson**[2] **Xiaonan Huang**[1]
[1]University of Michigan, Ann Arbor [2]Carnegie Mellon University [3]New York University

ABSTRACT

Depth ambiguity is a fundamental challenge in spatial scene understanding, especially in transparent scenes where single-depth estimates fail to capture full 3D structure. Existing models, limited to deterministic predictions, overlook real-world multi-layer depth. To address this, we introduce a paradigm shift from single-prediction to multi-hypothesis spatial foundation models. We first present `MD-3k`, a benchmark exposing depth biases in expert and foundational models through multi-layer spatial relationship labels and new metrics. To resolve depth ambiguity, we propose Laplacian Visual Prompting (LVP), a training-free spectral prompting technique that extracts hidden depth from pre-trained models via Laplacian-transformed RGB inputs. By integrating LVP-inferred depth with standard RGB-based estimates, our approach elicits multi-layer depth without model retraining. Extensive experiments validate the effectiveness of LVP in zero-shot multi-layer depth estimation, unlocking more robust and comprehensive geometry-conditioned visual generation, 3D-grounded spatial reasoning, and temporally consistent video-level depth inference. Our benchmark and code will be available at `https://github.com/Xiaohao-Xu/Ambiguity-in-Space`.

1 INTRODUCTION

Spatial understanding, the ability to derive structured 3D representations from sensory data, is fundamental to visual intelligence and autonomous systems. Despite progress in both physical sensors and monocular depth estimation models (Ranftl et al., 2022; Birkl et al., 2023; Bhat et al., 2023; Yin et al., 2023a; Guizilini et al., 2023; Li et al., 2024; Ke et al., 2024; Yang et al., 2024a; Gui et al., 2024; Fu et al., 2024; Piccinelli et al., 2024; Yin et al., 2023b; Yang et al., 2024b) (see Fig. 1a&b), a key challenge persists: **biased 3D spatial understanding under depth ambiguity**.

In real-world 3D scenes, factors such as transparency (see Fig.1c) break the assumption that each pixel corresponds to a unique depth value. For example, objects viewed through transparent surfaces like glass exhibit a range of plausible depths rather than a single fixed value. While state-of-the-art depth foundation models (Yang et al., 2024b;a) generalize well in unambiguous scenarios, they typically output only a single depth estimate, thereby ignoring inherent depth ambiguity. This limitation results in **biased, incomplete 3D representations** that undermine both generalization and reliability, especially in safety-critical applications requiring robust spatial reasoning.

To this end, we advocate a paradigm shift from single-prediction to Multi-Hypothesis Spatial Foundation Models (MH-SFMs). We posit that **true spatial intelligence demands explicitly modeling and resolving ambiguity** rather than forcing a biased single-depth output. To address this, we propose a unified framework that enables multi-layer depth estimation from a monocular image via a single, domain-agnostic foundation model (see Fig. 1d).

To enable rigorous study of multi-layer spatial relationships under depth ambiguity, we introduce `MD-3k`, a benchmark featuring explicit labels for multilayer spatial relationships that goes beyond traditional single-depth metrics. Our analysis reveals that existing models exhibit significant depth biases under standard RGB input—some favoring nearer surfaces, others preferring farther ones (see Fig. 2a)—highlighting the limitations of the conventional single-depth prediction paradigm.

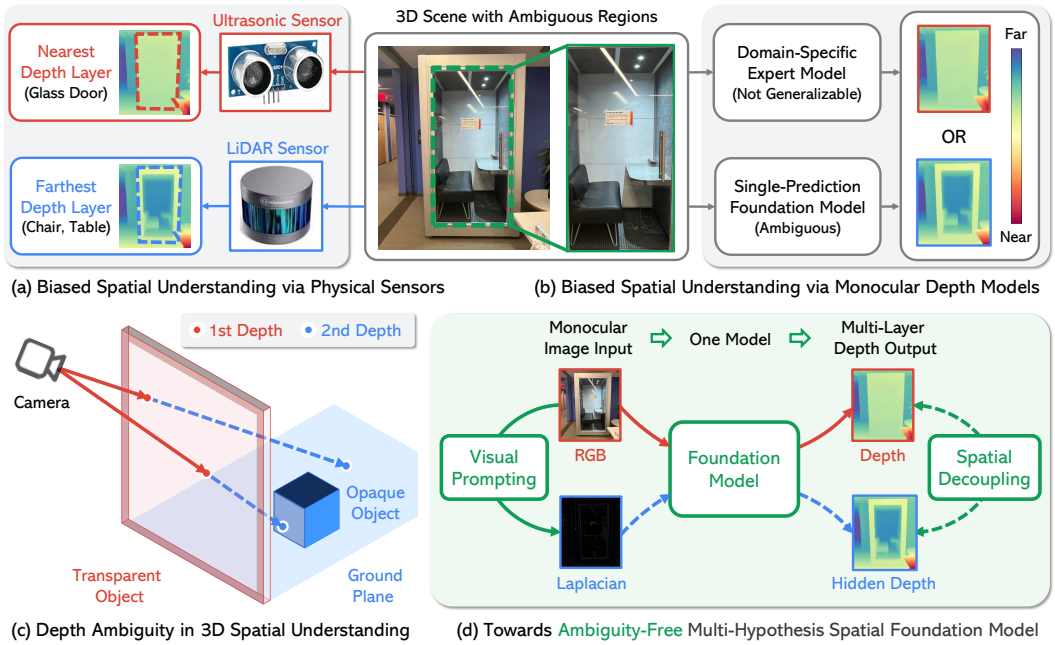

Figure 1: **Motivation.** 3D spatial understanding, powered by (**a**) sensors and (**b**) algorithms, has been confined to a biased ***single-layer*** representation of depth. (**c**) Existing methods collapse when faced with the true complexity of 3D, particularly in ambiguous scenes like those with transparency. (**d**) We propose Laplacian Visual Prompting (LVP) to transcend this limitation, granting *Spatial Foundation Models* the ability to derive ***multi-hypothesis*** depth, unlocking ambiguity-free spatial understanding.

Next, we introduce Laplacian Visual Prompting (LVP), a training-free spectral prompting technique for 3D spatial decoupling. LVP draws inspiration from prompting techniques in NLP (Wei et al., 2022) and visual prompting (Bar et al., 2022; Hojel et al., 2025; Bai et al., 2024). At its core, LVP applies the discrete Laplacian operator, a fundamental second-order difference operator, to the RGB image input. This operation generates high-frequency visual prompts that highlight regions of rapid intensity change, effectively exposing latent spatial knowledge within pre-trained depth models. Integrating depth maps from LVP and RGB inputs enables multi-hypothesis depth estimation without retraining, revealing pre-trained models' latent ability to disentangle multi-layered 3D structures. We demonstrate LVP's effectiveness on the `MD-3k` benchmark, showing that it uncovers hidden depth (see Fig. 2b) and mitigates inherent depth biases. Further analysis using LVP explores the scaling laws of spatial understanding under ambiguous and non-ambiguous scenes, and identifies key challenges in resolving multi-layer spatial relationships.

Finally, we demonstrate the practical benefits of LVP's multi-hypothesis depth estimation, enabling flexible geometry-conditioned visual generation (Zhang et al., 2023), including realistic 3D re-synthesis of transparent structures for ambiguous scenes, consistent multi-layer depth estimation in real-world videos, and robust 3D spatial reasoning in multi-modal LLMs. These results highlight the potential of LVP in advancing spatial intelligence.

Our main contributions are: **1**) We rethink spatial ambiguity in real-world 3D scenes and reformulate domain-agnostic, (*i.e.*, foundational) depth estimation as multi-hypothesis inference. **2**) We introduce `MD-3k`, a new benchmark to evaluate multilayer spatial understanding and model biases. **3**) analyze existing models across diverse architecture, training schema, and model size on `MD-3k` and reveal different depth biases under ambiguity. **4**) We propose Laplacian Visual Prompting (LVP), a training-free prompting method, to facilitate multi-hypothesis depth estimation from pre-trained models. **5**) We validate LVP's effectiveness in revealing multi-layer depth and depth bias control.

2 Related Work

Monocular depth estimation (MDE). MDE has evolved from early domain-specific depth estimation (Eigen et al., 2014; Fu et al., 2018; Bhat et al., 2021), constrained by dataset-specific training (*e.g.*,

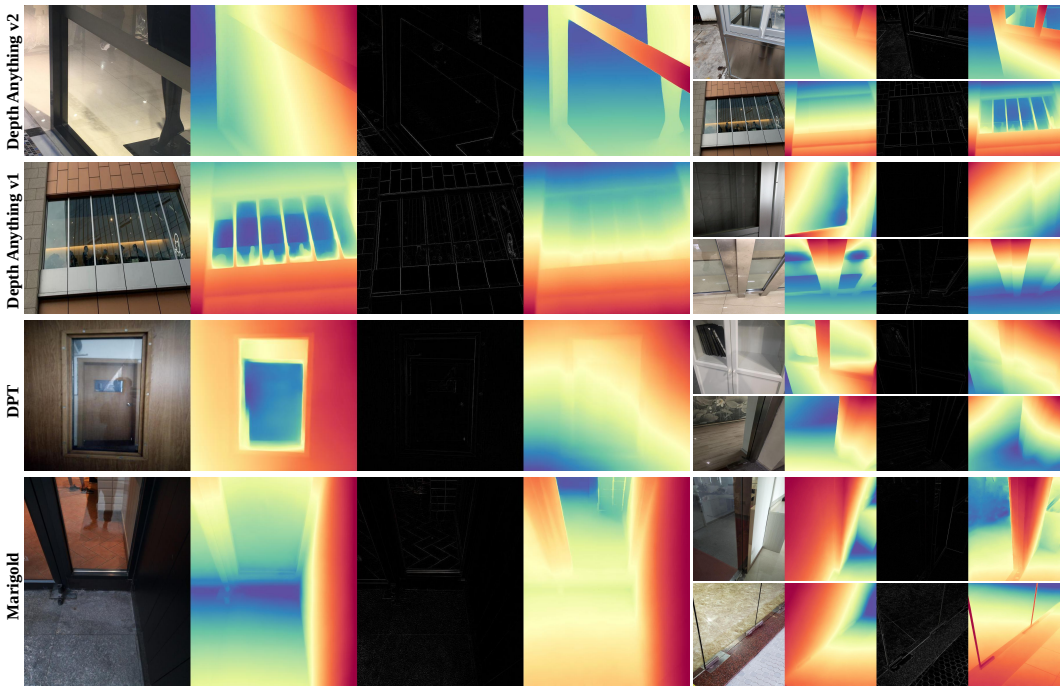

Figure 2: **Unlocking *hidden depth* with Laplacian Visual Prompting across diverse baselines (Yang et al., 2024a;b; Ke et al., 2024; Ranftl et al., 2021).** Each case includes the RGB input, estimated depth from RGB, Laplacian input, estimated *hidden depth* from Laplacian, and an enhanced Laplacian. Notice that depth maps from RGB and LVP **both capture *plausible hypotheses***: one for the ***transparent*** surface (glass) and another for the ***opaque*** object behind it.

NYU (Silberman et al., 2012), KITTI (Geiger et al., 2013)), to more generalizable domain-agnostic approaches, pushing forward the frontier towards generic depth foundation models. Recent methods exploit Stable Diffusion (Rombach et al., 2022) for fine-grained depth prediction (Ke et al., 2024; Gui et al., 2024; Fu et al., 2024). MiDaS (Ranftl et al., 2022; 2021; Birkl et al., 2023) and Metric3D (Yin et al., 2023a) rely on labeled data, while Depth Anything V1 (Yang et al., 2024a) and V2 (Yang et al., 2024b) enhance robustness through large-scale and pseudo-labeled training. Despite these advances, ***existing monocular depth foundation models estimate only single-layer depth, struggling with multi-layer spatial ambiguities.*** To address this, we redefine depth estimation in a domain-agnostic setting as a multi-hypothesis problem, using Laplacian Visual Prompting to disentangle depth layers in ambiguous visual contexts.

Visual prompting (VP). Inspired by prompt-based adaptation in NLP (Brown et al., 2020), VP (Bahng et al., 2022; Bai et al., 2024) enables pre-trained vision models to be adapted via input-space manipulation. VP has been successfully applied to vision-language models (Bahng et al., 2022; Singha et al., 2023; Wasim et al., 2023), with further improvements through joint text-visual optimization (Khattak et al., 2023; Wang et al., 2024). In addition, VP has been explored for black-box model adaptation (Tsai et al., 2020), cross-domain transfer (Chen et al., 2021; Neekhara et al., 2022), and adversarial robustness (Chen et al., 2023). While VP research has primarily focused on semantic understanding tasks, ***its potential for 3D spatial decoupling and comprehension remains largely unexplored.*** To address this gap, we introduce Laplacian Visual Prompting, which facilitates training-free spatial 3D decoupling through multi-hypothesis depth estimation.

3 Multi-Hypothesis Depth Estimation

Monocular depth estimation in complex 3D scenes is a multi-hypothesis inference problem, especially in transparent scenarios with multiple plausible depths.[1] To address this, we propose: 1) the `MD-3k` benchmark, which includes multi-layer spatial relationship labels, 2) new metrics for quantifying

[1]We consider ambiguous scenes with **two** visible depth layers, leaving more depth layers for future work.

single-layer depth estimation bias and multi-layer spatial relationship accuracy, and 3) a training-free spectral prompting method, *i.e.*, Laplacian Visual Prompting, to estimate multi-layer depth.

3.1 MD-3K BENCHMARK: QUANTIFYING MULTI-LAYER SPATIAL RELATIONSHIPS

The MD-3k benchmark quantifies spatial bias in depth estimation and evaluates multi-layer depth in ambiguous scenarios, providing an empirical foundation for assessing layered 3D understanding.

Benchmark construction. MD-3k consists of 3,161 RGB images sourced from the GDD dataset (Mei et al., 2020), selected for depth ambiguity, such as transparency. Following previous spatial relationship benchmarks for non-ambiguous scenes (*e.g.*, DIW (Chen et al., 2016) and DA-2k (Yang et al., 2024b)), we randomly sample a sparse point pair within the ambiguous region for each image. Expert annotators assigned pairwise depth order labels to points both on and behind transparent surfaces, generating two annotation layers. As shown in Fig. 3, each sample includes an RGB image, segmentation masks, and two types of spatial relationship labels (*near* and *far*) for point pairs representing multi-layer depths. Annotation accuracy was rigorously validated through multi-round expert review.

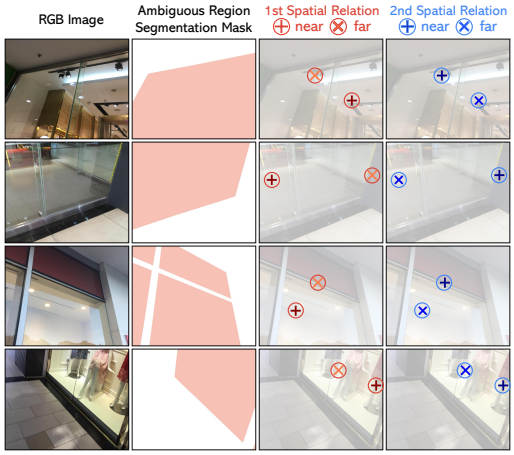

Figure 3: **MD-3k benchmark for evaluating multi-layer spatial relationships.** Example images feature annotated ambiguous region masks and sparse point pairs with multi-layer spatial labels. The first and second spatial relation columns show ground truth near/far annotations. The top three rows depict ***reverse*** relationships, while the bottom row shows a ***same*** relationship between layers.

Benchmark statistics. The full MD-3k dataset, referred to as ***overall***, is divided into two subsets with different multi-layer spatial relationships for fine-grained analysis: **1**) ***Same*** subset (1,783 point pairs): Consistent multi-layer relative depth ordering for each point pair; **2**) ***Reverse*** subset (1,378 point pairs): Reversed multi-layer relative depth ordering for each point pair. These subsets facilitate the evaluation of depth estimation models under varying conditions of multi-layer spatial ambiguity and relative depth consistency.

Fig. 4 summarizes statistics of ambiguous regions in the MD-3k benchmark. The left panel shows a histogram of the ambiguous-to-total area ratio per sample, capturing diverse ambiguity levels from minimal to near-total scene ambiguity. The right panel's heatmap indicates a balanced spatial distribution with a slight center bias, resembling a Gaussian pattern that reflects natural scene compositions while minimizing regional biases.

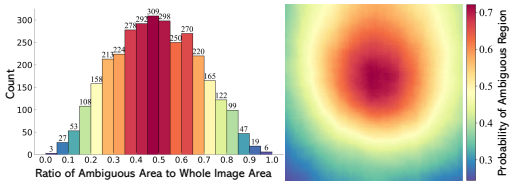

Figure 4: **Statistics of ambiguous regions in the MD-3k benchmark.** Ratio of ambiguous regions to the whole image (**Left**) and spatial distribution of ambiguity regions (**Right**)

3.2 METRICS: QUANTIFYING DEPTH BIAS AND MULTI-LAYER DEPTH ACCURACY

Spatial Relationship Accuracy (SRA(i)). SRA(i) measures the fraction of point pairs $\mathcal{P}$ with correct relative depth ordering for each depth layer $i \in \{1, 2\}$:

$$\text{SRA}(i) = \frac{1}{|\mathcal{P}|} \sum_{(P_1,P_2)\in\mathcal{P}} \mathbb{I}\Big(\text{sign}(\hat{d}_1^{(i)} - \hat{d}_2^{(i)}) = \text{sign}(d_1^{(i)} - d_2^{(i)})\Big), \tag{1}$$

where $\hat{d}_j^{(i)}$ and $d_j^{(i)}$ represent the predicted and ground truth depths at point P_j for layer i, respectively.

Depth Layer Preference ($\alpha(f_\theta)$). It quantifies the bias of a depth model f_θ towards one of the layers for layered scenes in its predictions. It is computed as the difference in SRA across layers:

$$\alpha(f_\theta) = \text{SRA}(2) - \text{SRA}(1). \tag{2}$$

A positive value ($\alpha(f_\theta) > 0$) indicates a preference for the second layer, while a negative value ($\alpha(f_\theta) < 0$) indicates a preference for the first layer. A higher absolute value signifies a stronger bias.

Multi-Layer Spatial Relationship Accuracy (ML-SRA). ML-SRA measures the proportion of point pairs where the predicted relative depth ordering is correct in both layers simultaneously:

$$\text{ML-SRA} = \frac{1}{|\mathcal{P}|} \sum_{(P_1,P_2)\in\mathcal{P}} \mathbb{I}\left(\bigwedge_{k=1}^{2} \text{sign}(\hat{d}_1^{(k)} - \hat{d}_2^{(k)}) = \text{sign}(d_1^{(k)} - d_2^{(k)}) \right). \tag{3}$$

3.3 Laplacian Visual Prompting for Multi-Layer Depth Decoupling

As shown in Fig. 5, we propose *Laplacian Visual Prompting* (LVP), a visual prompting technique designed to decouple multi-layer depth estimation by leveraging spectral components to resolve depth ambiguities in 3D scenes. LVP does not require retraining the depth model; instead, it employs a pre-trained monocular depth estimator to generate multiple depth hypotheses from a single RGB image. We posit that the latent depth distributions revealed by LVP can enhance depth estimation accuracy in scenarios with inherent depth ambiguity.

Probabilistic modeling of multi-hypotheses depth. To address depth ambiguity in monocular images, we propose a probabilistic model that predicts an ordered set of depth hypotheses, $\{\mathcal{D}_1, \mathcal{D}_2\}$, conditioned on the input image $\mathcal{I}$. To capture the relative depth ordering, we introduce a binary latent variable $\mathcal{O} \in \{0, 1\}$, where $\mathcal{O} = 1$ indicates that $\mathcal{D}_1 \prec \mathcal{D}_2$ (*i.e.*, $\mathcal{D}_1$ is closer than $\mathcal{D}_2$) and $\mathcal{O} = 0$ denotes that $\mathcal{D}_2 \prec \mathcal{D}_1$ (*i.e.*, $\mathcal{D}_2$ is closer than $\mathcal{D}_1$).

Rather than marginalizing over all possible orderings, we directly predict the ordered pair $(\mathcal{D}_1, \mathcal{D}_2)$ based on the sampled ordering $\mathcal{O}$:

$$p(\mathcal{D}_1, \mathcal{D}_2 \mid \mathcal{I}) = p(\mathcal{D}_1, \mathcal{D}_2 \mid \mathcal{O}, \mathcal{I})\, p(\mathcal{O} \mid \mathcal{I}). \tag{4}$$

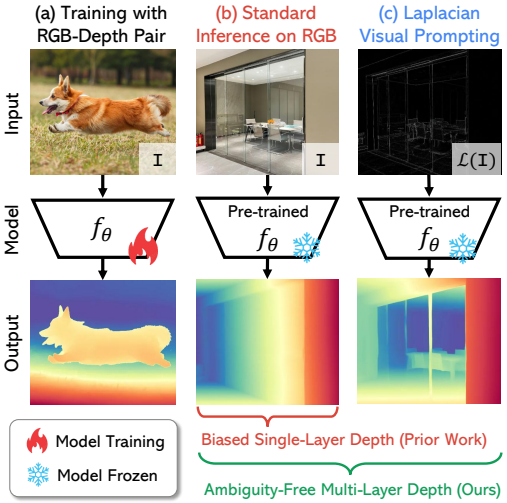

Figure 5: **Multi-layer depth with Laplacian Visual Prompting (LVP).** (**a**) Paired RGB-depth training of a domain-specific or domain-agnostic depth estimation model. (**b**) Standard inference via RGB input: single-layer depth on transparent glass. (**c**) Model inference via LVP: hidden depth revealing occluded objects, such as tables and chairs, behind the glass.

Assuming that $\mathcal{D}_1$ and $\mathcal{D}_2$ are independently estimated from $\mathcal{I}$ and that the single-layer depth prediction model is agnostic to the ordering $\mathcal{O}$, we derive:

$$p(\mathcal{D}_1, \mathcal{D}_2 \mid \mathcal{O}, \mathcal{I}) \propto p(\mathcal{D}_1 \mid \mathcal{I})\, p(\mathcal{D}_2 \mid \mathcal{I}). \tag{5}$$

Substituting into Eq. equation 4 yields:

$$p(\mathcal{D}_1, \mathcal{D}_2 \mid \mathcal{I}) \propto p(\mathcal{D}_1 \mid \mathcal{I})\, p(\mathcal{D}_2 \mid \mathcal{I})\, p(\mathcal{O} \mid \mathcal{I}), \tag{6}$$

where $p(\mathcal{D}_1 \mid \mathcal{I})$ and $p(\mathcal{D}_2 \mid \mathcal{I})$ represent the marginal likelihoods of the depth estimates for the two layers, and $p(\mathcal{O} \mid \mathcal{I})$ encodes the probability of the relative depth ordering. The relative ordering can be determined from the sign of layer preference $\alpha(f_\theta)$, as defined in Eq. equation 2.

Laplacian transformation for depth disambiguation. Monocular depth estimation often struggles with discontinuities and transparent surfaces, which leads to depth ambiguity. To mitigate this bias, we introduce Laplacian Visual Prompting, a spectral prompting strategy that uses the Laplacian operator to enhance the input image by emphasizing high-frequency details like object boundaries and edges. The Laplacian, a second-order derivative, effectively acts as a high-pass filter in the spatial domain. The 2D spatial Laplacian operator is defined as:

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \tag{7}$$

For an RGB image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, the Laplacian transformation is applied channel-wise:

$$\mathcal{L}(\mathcal{I}) = (\Delta\mathcal{I}_R; \Delta\mathcal{I}_G; \Delta\mathcal{I}_B), \tag{8}$$

where $\Delta\mathcal{I}_R$, $\Delta\mathcal{I}_G$, and $\Delta\mathcal{I}_B$ are the Laplacian-transformed red, green, and blue channels.

In discrete form, the Laplacian operator is approximated via a second-order finite difference scheme using a 3×3 convolution kernel:

$$\mathcal{M}_{\mathcal{L}} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{9}$$

Multi-hypothesis depth estimation. We apply a pre-trained monocular depth model f_θ to the original monocular RGB image input and its Laplacian-transformed version separately, which generates complementary depth hypotheses:

$$\mathcal{D}_1 = f_\theta(\mathcal{I}), \quad \mathcal{D}_2 = f_\theta(\mathcal{L}(\mathcal{I})). \tag{10}$$

These two independent depth predictions, $\mathcal{D}_1$ and $\mathcal{D}_2$, are combined with the latent ordering probability $p(\mathcal{O} \mid \mathcal{I})$ as described in Eq. equation 4. This formulation addresses ambiguity by explicitly representing multiple depth hypotheses.

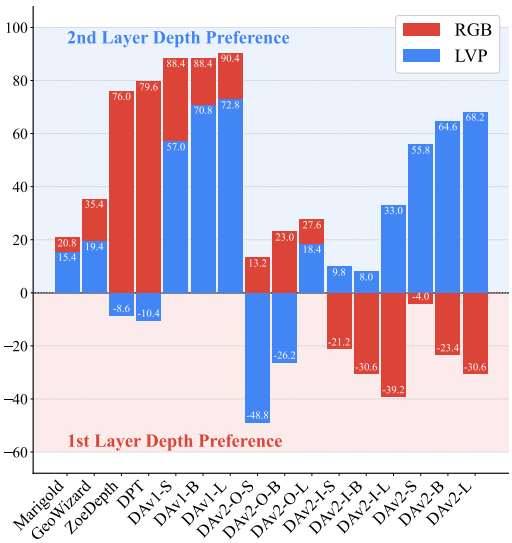

Figure 6: **Depth Layer Preference** $\alpha(f_\theta)$ **[%] under RGB and LVP Inputs on `MD-3k` (*Reverse*).** This figure highlights that the heterogeneous biases of standard RGB and LVP inputs significantly influence model preference, shifting it between the first and second annotated depth layers for certain models. This demonstrates how input modality can alter depth layer bias.

4 Experiments

To address the challenges of biased spatial understanding and unlock the potential of multi-hypothesis depth estimation, we raise the following fundamental questions: 1) **Depth bias** (Sec.4.1): In ambiguous scenes, what depth layer biases do existing models exhibit? 2) **LVP enhancement** (Sec.4.2): Can LVP effectively enhance multi-layer spatial understanding? 3) **Scaling laws** (Sec.4.3): How does model scale influence LVP-enhanced spatial understanding? 4) **Practical benefits** (Sec.4.4): What are the practical advantages of LVP-driven multi-hypothesis depth? 5) **LVP design** (Sec.4.5): How do LVP's design choices impact performance under ambiguity?

Baseline models. We assess bias and evaluate the effectiveness of our proposed multi-hypothesis depth estimation method via LVP across diverse baseline models, including Depth Anything V1/V2 (DAv1/2-S,B,L with ViT backbones (Yang et al., 2024a;b)). These models include general (DAv1/2), indoor (DAv2-I), and outdoor (DAv2-O) fine-tuned variants. Additional models include the discriminative models DPT (Ranftl et al., 2021) and ZoeDepth (Bhat et al., 2023), as well as the generative models Marigold (Ke et al., 2024) and GeoWizard (Fu et al., 2024).

4.1 Probing Single-Layer Depth Prediction Bias

We first analyze depth layer preference bias, which is defined in Eq. (2), for baseline models, revealing inherent biases in predicting closer or farther surfaces in ambiguous regions using standard RGB input. We then explore whether LVP can alter these biases by introducing complementary depth hypotheses to enrich RGB predictions.

Heterogeneous depth layer bias under standard RGB input. In Fig. 6, we observe significant heterogeneity in depth layer prediction preferences across models. Some models (*e.g.*, DAv2, DAV2-I) exhibit a bias towards the first depth layer, *i.e.*, $\alpha(f_\theta) < 0$, while others favor the second depth layer. In addition, models with the same architecture fine-tuned on different datasets (indoor/outdoor) can exhibit opposing depth biases, *e.g.*, DAv2-I and DAv2-O. This highlights how training data can hardwire assumptions about scene structure.

LVP modulates depth prediction preference. Comparing RGB and LVP results in Fig. 6 reveals that LVP effectively ***reverses*** or ***attenuates*** depth preferences across all baseline models. The pronounced

impact of LVP on depth preference modification, particularly in models like DAv2, suggests its ability to unlock latent representations. This reveals previously suppressed depth layers and fundamentally reshapes the model's depth interpretation.

True depth ambiguity exposes depth biases and remains challenging to tackle. True depth ambiguity, exemplified by scenes with ***reverse*** multi-layer spatial relationships (see Fig. 6), critically reveals depth layer biases. In these ambiguous scenes, performance becomes inconsistent (*i.e.*, large $|\alpha(f_\theta)|$) across RGB and LVP inputs, highlighting the difficulty in resolving conflicting spatial cues. Conversely, in non-ambiguous scenes with ***same*** multi-layer relationships (see Fig. 7), models achieve consistently high performance, exceeding 85% SRA under RGB input. This robustness is further supported by the small performance gap between RGB and LVP inputs on the non-ambiguous `DA-2k` benchmark, and the comparable RGB performance observed between `MD-3k` (*same*) and `DA-2k`. Thus, ambiguous scenes with ***reverse*** multi-layer relationships serve as a crucial diagnostic tool, effectively exposing the inherent depth biases and limitations of current depth baseline models.

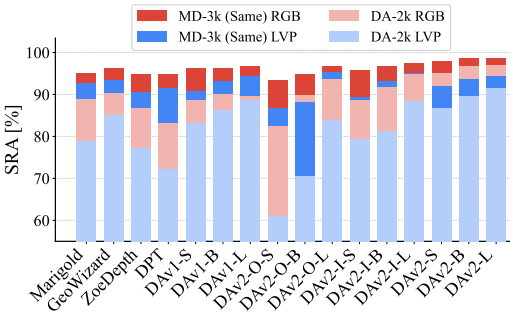

Figure 7: **High Spatial Relationship Accuracy (SRA) [%] under RGB and LVP inputs on the *same* subset of `MD-3k` and `DA-2k` (Yang et al., 2024b)** (non-ambiguous dataset for reference).

4.2 Multi-Layer Spatial Relationship Accuracy

Fig. 8 shows the Multi-Layer Spatial Relationship Accuracy (ML-SRA) achieved by our multi-hypothesis depth estimation method, which combines depth estimates from both RGB images and LVP inputs. This evaluation demonstrates the effectiveness of Laplacian Visual Prompting in generating complementary depth hypotheses beyond RGB-based inference, enabling ambiguity-free spatial understanding across diverse baselines.

Latent multi-layer knowledge suggests potential for MH-SFMs, unlocked by LVP. Despite being trained on single-layer depth data, some models implicitly capture multi-layer spatial relations. For example, DAv2, ZoeDepth, and DPT, when prompted with LVP, achieve non-trivial ML-SRA scores in challenging reverse spatial relationships (see Fig. 8b). This demonstrates that LVP effectively elicits this latent knowledge, suggesting that these models have the potential to be adapted into MH-SFMs, capable of representing and reasoning about multiple depth hypotheses. The fact that LVP is able to unlock this hidden potential highlights its significance as a key enabler for multi-hypothesis depth estimation.

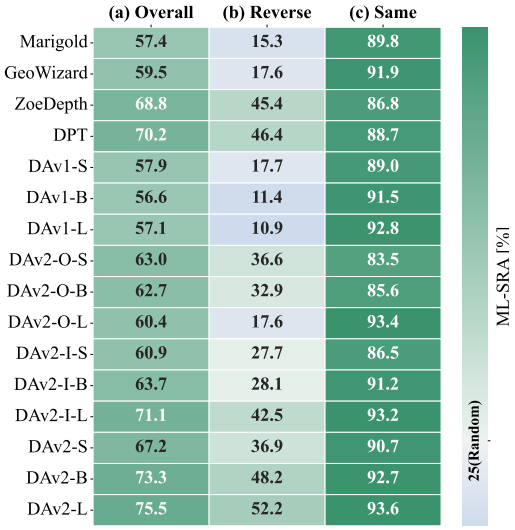

Figure 8: **Multi-Layer Spatial Relationship Accuracy (ML-SRA) [%] of our LVP-empowered multi-layer depth on `MD-3k`.** Effective performance gains of LVP-derived multi-depth over random guess (25%) are highlighted in green boxes.

Challenges in *reverse* relationships highlight the need for explicit ambiguity modeling, even with LVP. Accurately resolving depth ambiguity in *reverse* multi-layer spatial relationships remains challenging, even when using LVP to generate multi-layer depth estimates. The performance gap between *same* and *reverse* relationships highlights the difficulty of handling conflicting spatial cues and the limitations of relying solely on implicit priors, even when augmented by LVP. The reduced ML-SRA of domain-finetuned DAv2 models further suggests that optimizing for single-domain performance can hinder generalization to multi-layer scenes, reinforcing the need for models that can explicitly model and resolve ambiguity, rather than relying on domain-specific heuristics, even when combined with LVP-based prompting.

4.3 SCALING LAWS OF SPATIAL UNDERSTANDING

Developing generalist foundation models requires understanding performance scaling (Fan et al., 2024; Bai et al., 2024). We investigate how model scale impacts depth layer bias, multi-layer depth estimation in ambiguous scenes, and single-layer depth estimation in non-ambiguous scenes using Laplacian Visual Prompting (LVP), providing insights for building more reliable and trustworthy large-scale spatial foundation models.

To provide a high-level overview of these scaling behaviors, Fig. 9 summarizes the key performance trends observed in both ambiguous and non-ambiguous scenes as model scale increases. As depicted, **the impact of model scaling on spatial understanding is nuanced and context-dependent**, ***echoing a concurrent work on multi-modal alignment*** (Tjandrasuwita et al., 2025). Specifically, in ambiguous scenes, we observe divergent performance scaling depending on whether the model exhibits ***converged*** or ***diverged*** depth bias under RGB and LVP inputs. This divergence suggests that in scenarios with high spatial ambiguity (akin to high *uniqueness* in multi-modal data (Tjandrasuwita et al., 2025)), simply increasing model scale does not uniformly translate to improved performance. Instead, the *nature* of representation learning, specifically the depth bias, becomes a critical factor. Conversely, in non-ambiguous scenes, a more consistent pattern of generalization improvement with scale emerges, aligning with the expected benefits of larger model capacity in less challenging scenarios where redundancy is higher and ambiguity is lower.

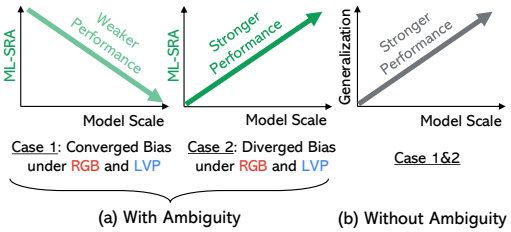

Figure 9: **Scaling laws of spatial understanding: performance trends in ambiguous vs. non-ambiguous scenes.** (**a**) In ambiguous scenes, converged depth bias (Case 1) leads to weaker performance with scale, while diverged bias (Case 2) yields stronger performance. (**b**) In non-ambiguous scenes, performance consistently improves with model scale, showing stronger generalization to the LVP input.

Model scale amplifies depth layer preference bias under RGB input. As shown in the left panel of Fig. 10, larger models tend to exhibit a stronger preference for certain depth layers under RGB input in ambiguous scenes with ***reverse*** multi-layer spatial relationships. DAv1 and DAv2-O models demonstrate a growing preference for the second depth layer while DAv2 and DAv2-I models demonstrating an increasing preference for the first layer.

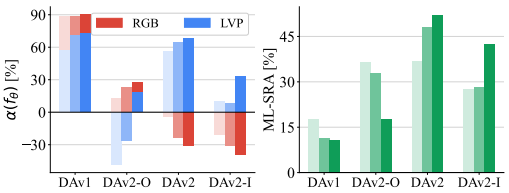

Figure 10: **Performance in ambiguous scenes (`MD-3k` *reverse* subset) as model scale increases. Left:** Depth Layer Preference $\alpha(f_\theta)$ [%]. **Right:** Multi-Layer Spatial Relationship Accuracy (ML-SRA) [%]. Bars within each group represent small, base, and large model variants (left to right).

Divergent depth bias elicit stronger multi-layer depth prediction with scale. As shown in the right panel of Fig. 10, larger models can exhibit a stronger divergence in depth layer preference based on input modality (RGB vs. LVP) in ambiguous scenes. Some models (DAv1, DAv2-O) ***converge*** towards a consistent second-layer preference, reducing multi-layer accuracy. Others (DAv2, DAv2-I) show a ***divergence***, improving ML-SRA with larger model, suggesting more diverse latent representations that encode multiple depth hypotheses.

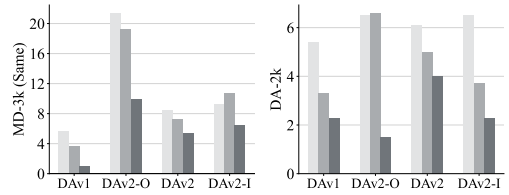

Figure 11: **Generalization to LVP input in non-ambiguous scenes as model scale increases**, measured by the SRA [%] gap between RGB and LVP inputs on the *same* subset of `MD-3k` (**Left**) and the non-ambiguous benchmark `DA-2k` (**Right**).

Enhanced generalization with increasing model scale in non-ambiguous scenes. Fig. 11 shows the performance gap between RGB and LVP inputs narrows as model scales. This improved generalization is observed in both multi-layer scenes with *same* spatial relationships of `MD-3k` and non-ambiguous scenes of `DA-2k`.

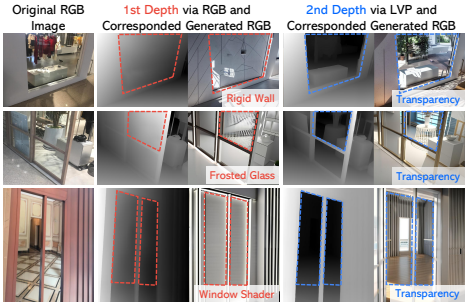

Figure 12: **Flexible 3D-conditioned visual generation with multi-layer depth.**

Figure 13: **Robust 3D spatial reasoning with LLM via multi-layer depth.**

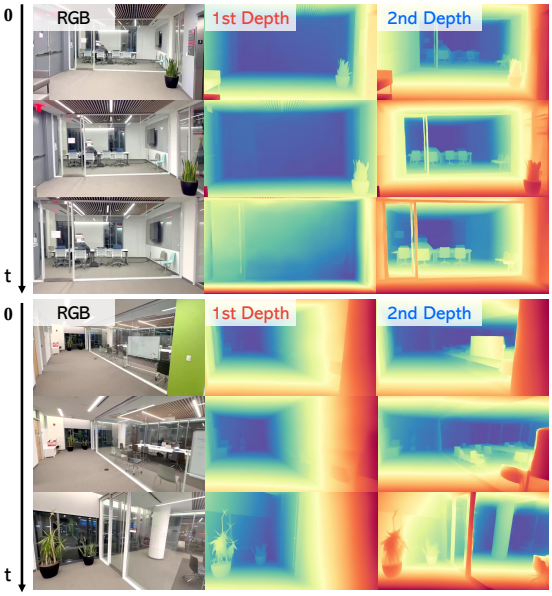

Figure 14: **Consistent multi-layer depth estimation on monocular RGB video sequence.**

4.4 Applications of Multi-Layer Depth

The multi-hypothesis depth predictions enabled by LVP enhance 3D-conditional image generation for ambiguous scenes. This capability supports the creation of complex environments, such as those featuring both transparent and opaque objects (*e.g.*, glass doors and windows), using geometry-conditioned ControlNet (Zhang et al., 2023) (see Fig.12). In addition, LVP boosts 3D spatial reasoning through a Multi-modal Large Language Model (LLM), exemplified by precise 3D-grounded human counting with the ChatGPT o3-mini model (see Fig.13). Furthermore, the multi-layer depth estimation via LVP also demonstrates robust consistency when applied to real-world videos (see Fig.14).

4.5 Ablation Study of LVP Design

Fig. 15 shows that ML-SRA performance is largely unaffected by Laplacian discretization (4-neighbor vs. 8-neighbor in LVP and LVP-2) and kernel sign (LVP-R with reversed convolution vs. LVP), with variations generally within $\pm 3\%$. While grayscale LVP (LVP-G) slightly reduces SRA compared to RGB LVP, the difference is minimal. These results highlight the crucial role of high-frequency information in 3D decoupling.

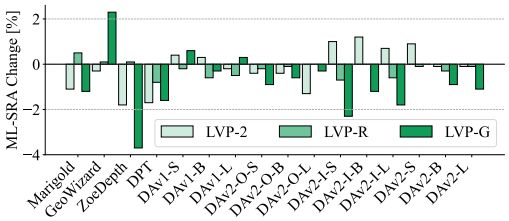

Figure 15: **Ablation study of LVP design**. Overall ML-SRA [%] change relative to the default LVP on `MD-3k` is shown.

5 Conclusion

We redefine domain-agnostic monocular spatial foundation models as inherently ambiguous, multi-hypothesis problems. To advance this, we introduce Laplacian Visual Prompting (LVP), a training-free technique for multi-layer depth estimation, and `MD-3k`, the first benchmark for evaluating multi-layer depth under ambiguity. Our analysis highlights significant biases in existing models, revealing the limitations of single-depth estimation. Experiments show that LVP modulates depth biases, enables comprehensive multi-layer estimation, and enhances downstream task robustness and flexibility.

Limitations & Future Work. Future work should broaden spatial understanding by exploring diverse multi-modal visual prompts, including learned spectral transformations. While `MD-3k` uses noisy real-world images, robustness to various noise types and artifacts needs evaluation. Developing benchmarks with real-world multi-layer depth annotations, potentially via sensor fusion, would improve performance assessment, though challenging. Finally, addressing diverse spatial ambiguities like reflection is crucial for reliable spatial foundation models.

REFERENCES

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022.

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *CVPR*, pp. 22861–22872, 2024.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *NeurIPS*, 35:25005–25017, 2022.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*, 2023.

Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

Aochuan Chen, Peter Lorenz, Yuguang Yao, Pin-Yu Chen, and Sijia Liu. Visual prompting for adversarial robustness. In *ICASSP*, pp. 1–5. IEEE, 2023.

Lingwei Chen, Yujie Fan, and Yanfang Ye. Adversarial reprogramming of pretrained neural networks for fraud detection. In *CIKM*, pp. 2935–2939, 2021.

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NeurIPS*, 2016.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *CVPR*, pp. 7382–7392, 2024.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *ECCV*, 2024.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013.

Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv:2403.13788*, 2024.

Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023.

Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors. In *ECCV*, pp. 257–273. Springer, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *CVPR*, 2023.

Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *CVPR*, 2024.

Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, pp. 3687–3696, 2020.

Paarth Neekhara, Shehzeen Hussain, Jinglong Du, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. Cross-modal adversarial reprogramming. In *WACV*, pp. 2427–2435, 2022.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Mainak Singha, Harsh Pal, Ankit Jha, and Biplab Banerjee. Ad-clip: Adapting domains in prompt space using clip. In *ICCV*, pp. 4355–4364, 2023.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. *arXiv preprint arXiv:2502.16282*, 2025.

Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. *ICML*, 2020.

Hao Wang, Fang Liu, Licheng Jiao, Jiahao Wang, Zehua Hao, Shuo Li, Lingling Li, Puhua Chen, and Xu Liu. Vilt-clip: Video and language tuning clip with multimodal prompt learning and scenario-guided optimization. In *AAAI*, volume 38, pp. 5390–5400, 2024.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *CVPR*, pp. 23034–23044, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35: 24824–24837, 2022.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024b.

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 *ICCV*, 2023a.

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 *ICCV*, pp. 9043–9053, 2023b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.