CONTEXTDEPTH: BOOSTING METRIC DEPTH ESTIMATION VIA RICHER CONTEXT

Anonymous authors
Paper under double-blind review

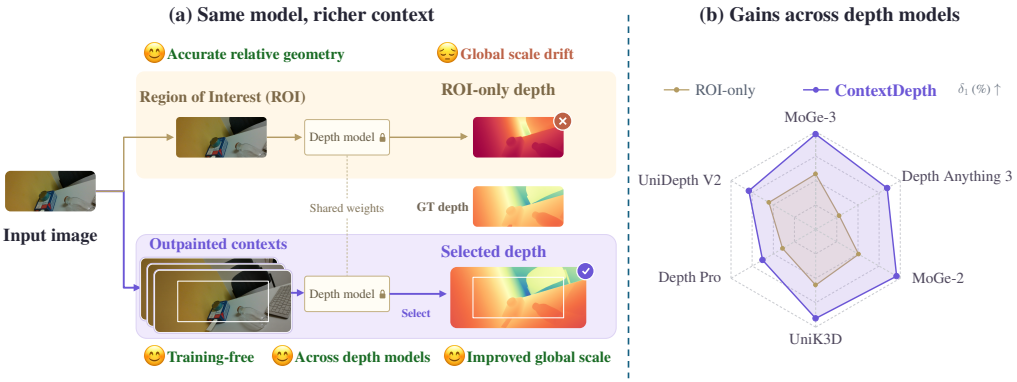

Figure 1: **Contextual depth reasoning with ContextDepth.** (a) Pretrained metric depth models recover sharp relative geometry but often suffer from global scale drift due to limited field-of-view. ContextDepth surrounds the query with outpainted contexts and uses a retrieved reference to select the scale-correct candidate, resolving scale ambiguity without updating model weights or modifying raw pixels. (b) Across six mainstream depth models, our training-free spatial context consistently boosts δ_1 metric accuracy.

ABSTRACT

Monocular metric depth estimation is inherently ill-posed: a single image provides only a keyhole view of the scene, causing models to misestimate absolute scale despite recovering sharp relative geometry. While in-context learning (ICL) steers models through inputs alone, existing visual ICL frameworks focus on instructing generalist models on *which* task to perform. We ask a different question: can context improve *how well* a fixed specialist model executes its task? We propose **ContextDepth**, a training-free spatial ICL framework that improves pretrained metric depth estimators without updating model weights or modifying raw input pixels. ContextDepth surrounds the query with diverse outpainted contexts to induce alternative scale hypotheses, then uses the directional bias from a retrieved depth reference to select the best candidate, leaving predicted depths unmodified. Across six mainstream depth models and six benchmarks, ContextDepth delivers consistent improvements with and without camera intrinsics, boosting δ_1 accuracy by up to 22.65 percentage points on the strongest baseline, Depth Anything 3. Crucially, guided selection remains robust under cross-dataset and out-of-domain reference banks where direct rescaling collapses, while in-domain banks yield the largest gains. These results establish visual context as an effective lever, unlocking substantial headroom as generative priors and reference banks scale.

1 INTRODUCTION

In-context learning (ICL) steers a pretrained model through its input alone: extended instructions placed in the prompt change the model's output without any parameter update [5]. Its effectiveness is well documented on LLMs, spanning few-shot demonstrations [1; 5] and chain-of-thought reasoning [35]. This principle was first established for text and has since been extended to vision, where a prompt of paired images tells a generalist model *which* task to perform [32; 40]. We ask a different question: can context improve *how well* a fixed vision model executes a task it already knows? We probe this on monocular metric depth estimation, which predicts the absolute distance, in meters, between the camera and every pixel of a single RGB image—an output underpinning robot

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.

navigation and manipulation, autonomous driving, and AR/VR [17; 37; 39]. The task offers a precise, quantitative measure of perceptual accuracy, so the benefit of added context is directly measurable.

Progress on metric depth, however, trails the relative domain. Feed-forward models recover accurate relative geometry with sharp boundaries [8; 30; 39], and diffusion-based estimators cast depth as a generative problem [6; 12; 21], yet accurate relative geometry alone does not determine absolute metric scale. Prior methods bridge the relative-to-metric gap via metric heads on relative backbones, canonical camera frames, or joint prediction of depth and intrinsics [3; 4; 10; 14; 22; 23; 31; 38]. However, these methods still require metric training supervision. The bottleneck is data: paired RGB-D captures are far costlier than RGB images [8]. Beyond data, the task poses two fundamental challenges. First, depth is ambiguous: a single image is consistent with infinitely many 3D scenes [15], making single-view recovery inherently one-to-many even with known camera intrinsics. Second, the camera frame offers only a keyhole view that cannot uniquely resolve overall spatial layout. The same local observation can thus match multiple scene configurations with different metric depths.

These observations motivate our goal: improving metric depth estimators without retraining, by enriching the visual context the observation leaves out.

Q1 — How can context be added to an image?
A: Outpaint the image, optionally with retrieved RGBD references of similar scenes, to guide selection among outpainting candidates: the former adds visual cues, the latter anchors metric scale.

Our answer is a spatial form of in-context learning for metric depth models, built on two complementary mechanisms. **1) Outpainting**: we surround the input image with different generated contexts while keeping its raw pixels fixed; the same depth model thus produces different depth estimates for identical input pixels. In effect, an outpainting model lends its scene priors to the depth model, reshaping how the image is interpreted. A generated context, however, can help or harm: outpainting may extend the scene plausibly yet geometrically inconsistently, misleading the scale estimate. This motivates the second mechanism. **2) Reference guidance**: we retrieve an independent image with known depth from a similar scene, and use it as scale evidence to select among the candidate contexts. Crucially, instead of *directly rescaling* predictions with reference error magnitude, we perform *guided selection*: reference error provides only a directional bias to select among candidate contexts, leaving predicted depth values unmodified. For instance, when depth is systematically overestimated in similar scenes, guided selection favors candidates yielding smaller depths.

Q2 — Do these contexts improve metric depth estimation?
A: Yes, on both counts. On Depth Anything 3 (DA3), outpainting alone lifts δ_1 by 17.3 points $\uparrow$, and reference-guided selection adds 5.4 more $\uparrow$, with model weights frozen.

We evaluate both mechanisms on 3,406 images across six benchmarks and six depth models, with and without query intrinsics (Sec. 4.1). Three conclusions emerge. **First**, gains stem from context content rather than the larger canvas: on the same expanded canvas, a single generated context beats all padding controls in 11 of 12 settings. **Second**, multiple candidates offer substantial headroom: the best-of-N oracle gains up to 10.62 δ_1 points $\uparrow$ (Fig. 3), motivating an effective selector. **Third**, reference-guided selection converts this headroom into real gains: up to 22.65 δ_1 points $\uparrow$ on DA3 [18] and consistent improvements across models (Tab. 1); it also stays robust under out-of-domain reference banks, where direct rescaling collapses (Tabs. 2 and 3).

Our main contributions are threefold:

- We propose *contextual depth reasoning*, a training-free paradigm that treats the input, rather than the weights, as the lever for improving metric depth estimation—extending ICL from steering *which* task a model performs to improving *how well* it performs.
- We instantiate this paradigm with two complementary mechanisms, outpainting-based context expansion and reference-guided scale evidence, along with a reference-guided selector that retains the gains while discarding harmful candidates.
- Extensive experiments across six mainstream models and six public benchmarks show consistent gains, while ablations establish when context helps, harms, or requires selection.

2 Related Work

2.1 Metric Depth Estimation

Monocular metric depth estimation recovers absolute 3D scene geometry from a single RGB image, where camera-dependent scale ambiguity and cross-scene generalization remain fundamental challenges. ZoeDepth [3] incorporates metric depth heads on relative-depth backbones, while Metric3D [38] normalizes camera views into a canonical space to resolve scale ambiguity across different focal lengths. The UniDepth family [22; 24] and UniK3D [23] jointly predict depth and camera intrinsics to accommodate diverse uncalibrated setups. Recent work has focused on improving scale fidelity and high-frequency geometric detail: Depth Pro [4] employs a multi-scale architecture for zero-shot metric depth, MoGe-2 [31] predicts metric point maps directly, MoGe-3 [14] refines geometric details through sparse volumetric representations, and Depth Anything 3 [18] unifies monocular and multi-view geometry via a depth–ray formulation. These methods primarily enhance performance through model architectures, camera modeling, or supervised training on metric datasets. In contrast, we propose a training-free inference framework that enriches spatial context and uses reference guidance to improve metric depth predictions on a frozen model.

2.2 Visual In-Context Learning

In-context learning steers pretrained models through input prompts alone without updating model parameters [5]. In computer vision, visual in-context learning provides paired visual examples to guide task execution at inference time. Visual Prompting [2] formulates in-context prediction as image inpainting on example grids, while Painter [32] and SegGPT [33] adopt unified output representations to enable example-driven multi-task prediction and segmentation. Example quality and composition also directly influence prediction accuracy: Zhang et al. [40] investigate demonstration retrieval and selection, while Condenser [29] and PromptHub [19] aggregate multiple visual prompts to combine complementary evidence. Beyond discriminative tasks, PromptDiffusion [34] and VisualCloze [16] adapt visual in-context learning to generative diffusion models, and SD-VICL [20] shows that off-the-shelf diffusion models possess in-context capabilities without fine-tuning. While prior works use in-context examples to specify *which* task a generalist vision or diffusion model should perform, we apply spatial in-context learning to a single-task specialist, probing whether generated spatial context and reference evidence can improve *how well* a fixed metric depth estimator performs.

3 Method

Our goal is to apply in-context learning (ICL) to monocular metric depth estimation: introducing visual context to improve metric accuracy without model finetuning (Fig. 2). We first construct spatial context around the input image (Sec. 3.1). Then, a reference image with known depth guides candidate selection with or without known query intrinsics (Sec. 3.2). See *Sup.Mat.*'s Sec. B for a latent-concept view and relative-depth analyses.

3.1 Spatial Context Construction

Given an RGB image x, an outpainting model, OneReward [9], generates visual context beyond its boundaries, with x centered on the expanded canvas (Fig. 2(a)). We refer to x as the region of interest (ROI) pixels or the query image. The generated surroundings form the spatial context for estimating the metric depth of the ROI pixels. The outpainting model may slightly alter pixels inside the ROI. We therefore paste the original pixels back, so that only the generated surroundings serve as context. Repeating this process yields N candidates $\hat{\mathcal{X}} = \{\hat{x}_i\}_{i=1}^{N}$, $i \in \{1, \ldots, N\}$.

Next, each candidate is processed independently by the same frozen depth model, and the prediction is cropped back to the ROI at its original resolution, giving $\hat{d}_i$. Over all candidates this yields the prediction set $\hat{\mathcal{D}} = \{\hat{d}_i\}_{i=1}^{N}$ and, when the model estimates intrinsics, the set $\hat{\mathcal{K}} = \{\hat{k}_i\}_{i=1}^{N}$.

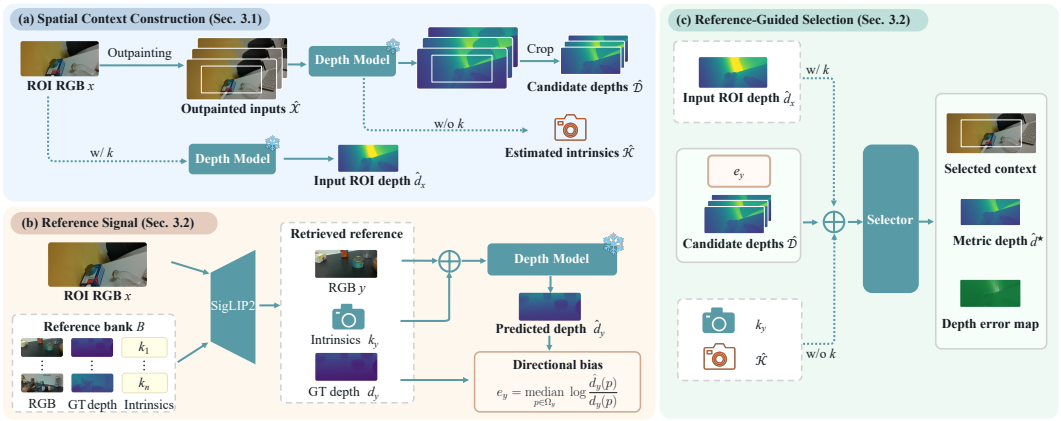

Figure 2: **Overview of ContextDepth.** **(a)** We generate multiple outpainted contexts while preserving the original ROI pixels, run the same frozen depth model on each expanded image, and crop the predictions back to the ROI. **(b)** We retrieve a reference image similar to the input and compare the model's prediction with the ground-truth reference depth map to obtain a directional bias. **(c)** The Selector uses the directional bias to guide selection and outputs the final metric depth within the ROI.

Algorithm 1 Reference-guided outpainting selection

Input: Candidates $\hat{\mathcal{X}} = \{\hat{x}_i\}_{i=1}^N$ with predictions $\hat{\mathcal{D}} = \{\hat{d}_i\}$; reference (y, d_y, k_y); intrinsics mode
Unknown mode: estimates $\hat{\mathcal{K}} = \{\hat{k}_i\}$; Known mode: query intrinsics k and baseline $\hat{d}_x$
$\mathcal{F}$: frozen depth model; δ: intrinsics distance (Eq. 2); med log: pixelwise median log ratio (Eqs. 1, 3, 5)
Output: Final query depth $\hat{d}^\star$

$\hat{d}_y \leftarrow \mathcal{F}(y;\, k_y)$
$e_y \leftarrow \underset{p\in\Omega_y}{\text{median}} \log \frac{\hat{d}_y(p)}{d_y(p)}$
if k unknown **then**
  $a \leftarrow \arg\min_i \delta(\hat{k}_i,\, k_y)$
  $\mathcal{E} \leftarrow \{i : e_y \cdot \Delta_i < 0\}$
  if $\mathcal{E} \neq \emptyset$ **then**
    $i^\star \leftarrow \arg\min_{i\in\mathcal{E}} \delta(\hat{k}_i,\, k_y)$
  else
    $i^\star \leftarrow a$
return $\hat{d}^\star \leftarrow \hat{d}_{i^\star}$
else ▷ k known
  $\mathcal{E} \leftarrow \{i : \Delta_i < 0\}$ **if** $e_y > 0$ **else** $\{i : \Delta_i \geq 0\}$
  $i^\star \leftarrow \arg\max_{i\in\mathcal{E}} |\Delta_i|$
  if $\mathcal{E} = \emptyset$ **then**
    return $\hat{d}^\star \leftarrow \hat{d}_x$

3.2 Reference-Guided Context Selection

3.2.1 Overview

For a query x, Sec. 3.1 yields the candidate outpaintings $\hat{\mathcal{X}}$ with depth predictions $\hat{\mathcal{D}}$. A reference retrieved from the reference bank then guides the choice of the final prediction $\hat{d}^\star$.

The reference bank stores RGB images with ground-truth depth (i.e., RGBD scans) and intrinsics. A frozen encoder SigLIP2 [27] retrieves the image visually most similar to the query, yielding the triplet (y, d_y, k_y); the bank's construction and composition are examined in Sec. 4.

The selection exploits the **model's scene-dependent bias**. As discussed in *Sup.Mat.*'s Sec. B, scale estimation is prior-driven, so the model errs in a direction that depends largely on the type of scene. Since the reference depicts a scene similar to the query, the model tends to err on it in the same direction as on the query. Comparing the model's prediction on the reference against its ground-truth depth therefore estimates this bias: whether the model tends to over- or under-estimate depth for scenes of this type. Selection then keeps the candidate whose scale change corrects this bias, i.e. whose implied scale shift moves the prediction toward the true scale.

The reference also supplies **camera calibration**, and how it does so depends on whether the query intrinsics are available. When k are unknown, the reference intrinsics k_y serve as the calibration for depth estimation. In summary, the selection favors candidates that correct the model bias and have intrinsics close to the reference; Algorithm 1 summarizes the entire selection procedure. Sec. 3.2.2 and Sec. 3.2.3 present the two branches in turn.

3.2.2 SELECTION WITH UNKNOWN QUERY INTRINSICS

We first run the same depth model on the reference image y with the known intrinsics k_y, obtaining $\hat{d}_y$, and compute the signed median log ratio of the prediction to the ground-truth depth d_y:

$$e_y \;=\; \underset{p\in\Omega_y}{\text{median}}\; \log \frac{\hat{d}_y(p)}{d_y(p)}. \tag{1}$$

Here, e_y is a *directional bias*: $e_y > 0$ means the model overestimates reference depth, and $e_y < 0$ means it underestimates. The ratio and logarithm are pixelwise, and $\Omega_y \subset \{1,\dots,H\}\times\{1,\dots,W\}$ collects the pixels where both depths are finite and positive; the same convention defines the valid sets in all subsequent equations. This sign sets the depth-change direction for filtering.

Each outpainting $\hat{x}_i \in \hat{\mathcal{X}}$ comes with a depth prediction $\hat{d}_i$ and model-estimated intrinsics $\hat{k}_i \in \hat{\mathcal{K}}$. The anchor $\hat{x}_a$ is the candidate whose estimated intrinsics are closest to k_y:

$$a \;=\; \arg\min_i\, \delta\Big(\hat{k}_i,\, k_y\Big), \quad \text{where} \quad \delta\Big(\hat{k}_i,\, k_y\Big) \;=\; \sum_{\alpha\in\{u,v\}} \Big(\log \frac{\hat{f}_i^{\alpha}}{f_y^{\alpha}}\Big)^2. \tag{2}$$

Here, u and v denote the horizontal and vertical image axes, respectively, and $\hat{d}_a$ is the baseline for comparison. The distance δ compares the model-reported focal components $\{\hat{f}_i^{\alpha}\}_{\alpha\in\{u,v\}}$ with the reference focal components $\{f_y^{\alpha}\}_{\alpha\in\{u,v\}}$, each normalized by its image dimension for comparability.

Each candidate's depth change from the anchor is the median of the pixelwise log-depth ratio over the common query pixels Ω_{ROI} where all N predictions are finite and positive:

$$\Delta_i \;=\; \underset{p\in\Omega_{\text{ROI}}}{\text{median}}\; \log \frac{\hat{d}_i(p)}{\hat{d}_a(p)}. \tag{3}$$

We retain candidates whose change has the opposite sign to e_y, i.e. the eligible set $\mathcal{E} = \{i : e_y \cdot \Delta_i < 0\}$, and among them select the one whose estimated intrinsics are closest to k_y:

$$i^\star \;=\; \begin{cases} \arg\min_{i\in\mathcal{E}}\, \delta\Big(\hat{k}_i,\, k_y\Big), & \text{if } \mathcal{E}\neq\emptyset, \\ a, & \text{otherwise.} \end{cases} \tag{4}$$

The selected prediction is returned as $\hat{d}^\star = \hat{d}_{i^\star}$ without rescaling.

3.2.3 SELECTION WITH KNOWN QUERY INTRINSICS

When the query intrinsics k are known, both the query x and each candidate $\hat{x}_i \in \hat{\mathcal{X}}$ are processed with this calibration, using the query-only prediction $\hat{d}_x$ as the baseline. The sign of e_y sets the required direction of change: candidates with $\Delta_i < 0$ qualify when $e_y > 0$, and those with $\Delta_i \geq 0$ qualify when $e_y \leq 0$. Among qualifying candidates, we select the one with the largest change:

$$i^\star \;=\; \begin{cases} \arg\max_{i:\,\Delta_i<0} \big|\Delta_i\big|, & \text{if } e_y > 0, \\ \arg\max_{i:\,\Delta_i\geq 0} \Delta_i, & \text{if } e_y \leq 0, \end{cases} \qquad \text{where} \qquad \Delta_i \;=\; \underset{p\in\Omega_x}{\text{median}}\; \log \frac{\hat{d}_i(p)}{\hat{d}_x(p)}. \tag{5}$$

The final prediction is $\hat{d}^\star = \hat{d}_{i^\star}$ if any candidate qualifies, and $\hat{d}^\star = \hat{d}_x$ otherwise.

4 EXPERIMENTS

4.1 SETTINGS

Datasets. We evaluate on 3,406 queries from six datasets: NYUv2 [26], KITTI [7], ETH3D [25], HAMMER [11], iBims-1 [13], and DIODE [28], using the preprocessed evaluation data of MoGe-2 [31]. For each image, the query ROI is a centered crop covering about two-thirds of its width and height; it is the input to outpainting and is fixed across all settings. We evaluate all predictions against ground-truth depth on valid pixels within this ROI.

Table 1: **Metric depth estimation with the mixed reference bank.** Cell shading indicates metric quality within each camera setting via a continuous red-to-green gradient. **Bold** and underlined entries mark the best and second-best results per model row; Mean Δ averages changes relative to Query-only.

Model	*w/o query intrinsics*						*w/ query intrinsics*					
	Query-only		Outpaint		Outpaint + Select		Query-only		Outpaint		Outpaint + Select	
	$\delta_1\uparrow$	AbsRel↓	$\delta_1\uparrow$	AbsRel↓	$\delta_1\uparrow$	AbsRel↓	$\delta_1\uparrow$	AbsRel↓	$\delta_1\uparrow$	AbsRel↓	$\delta_1\uparrow$	AbsRel↓
MoGe-3 (ViT-G)	62.67	0.2309	76.73	0.2248	**78.99**	**0.2038**	78.36	0.2368	78.84	0.2276	**81.21**	**0.1970**
Depth Anything 3	51.17	0.2361	68.47	0.2038	**73.82**	**0.1788**	73.99	0.2056	79.64	0.1724	**81.05**	**0.1555**
MoGe-2	60.32	0.2075	77.06	0.1970	**78.28**	**0.1814**	80.39	0.1954	82.02	0.1854	**82.83**	**0.1616**
UniK3D (ViT-L)	62.80	0.2625	75.24	0.2658	**76.43**	**0.2315**	76.26	0.2457	76.74	0.2349	**78.62**	**0.1850**
Depth Pro	55.62	0.2527	63.39	0.2483	**65.08**	**0.2331**	68.12	0.2248	67.75	0.2236	**70.56**	**0.1981**
UniDepth V2	62.07	0.2917	70.13	0.2847	**71.58**	**0.2608**	72.89	0.2777	72.71	0.2784	**75.32**	**0.2214**
Mean Δ	–	–	**+12.73**	**−0.0095**	**+14.92**	**−0.0320**	–	–	**+1.28**	**−0.0106**	**+3.26**	**−0.0446**

Reference Banks. The reference bank supplies annotated examples (RGB image, ground-truth depth, and intrinsics) to guide candidate selection. We evaluate three configurations: 1) the default *mixed bank* pools references from all six datasets (excluding the query itself and its scene/capture group to reduce the influence of near-duplicate content) so retrieval can benefit from in-domain examples; 2) the *cross-dataset bank* excludes references from the query dataset to test robustness when in-domain examples are unavailable; and 3) the *fixed DIODE bank* uses 771 DIODE images as references for all queries to test whether a single bank can serve diverse targets without per-dataset adaptation. To ensure fair comparison across reference banks (Tab. 2) and strategy ablations (Tab. 3), all columns and Query-only baselines are evaluated on the identical set of 2,635 non-DIODE queries.

Baselines. We evaluate six pretrained models for monocular metric depth estimation: Depth Anything 3 [18], MoGe-2 [31], Depth Pro [4], UniDepth V2 [24], MoGe-3 (ViT-G) [14], and UniK3D (ViT-L) [23]. We consider settings with and without query intrinsics. Except for Depth Anything 3, we use the largest public checkpoint of each model, which supports both settings with fixed weights. For Depth Anything 3, we use DA3Metric-Large with known query intrinsics and DA3Nested-Giant-Large without. All model weights remain frozen throughout. For each query, we generate $N = 3$ outpainting candidates using the OneReward model [9] under three random seeds, restore the original query pixels, and run depth inference and selection on an NVIDIA GeForce RTX 4090 GPU. With query intrinsics, all models use a 25% outpainting extent per ROI side; without intrinsics, we use 25% for MoGe-3, DA3, and MoGe-2, and 10% for UniDepth V2, UniK3D, and Depth Pro. Details are in *Sup.Mat.*'s Sec. C, including the rationale for these extents in *Sup.Mat.*'s Sec. D.1.

Metrics. For predicted depth $\hat{d}$ and query ground-truth depth d_x, we report threshold accuracy ($\delta_1\uparrow$, %), defined as the percentage of pixels satisfying $\max(\hat{d}/d_x, d_x/\hat{d}) < 1.25$, and absolute relative error (AbsRel↓), defined as the mean of $|\hat{d} - d_x|/d_x$. Specifically, δ_1 measures inlier coverage within a 25% error tolerance, whereas AbsRel penalizes continuous proportional errors across all pixels; reporting both guards against conclusions driven by a single error regime. Both metrics are computed over valid pixels within the fixed query ROI and averaged across queries without test-time scale or shift alignment; ablation and reference-bank analyses (Tabs. 2 and 3 and Fig. 3) report δ_1, with AbsRel counterparts in *Sup.Mat.*'s Sec. D.3.

4.2 Comparisons

Quantitative Results. Table 1 quantifies the two core contributions of our method across all six depth models: the effect of outpainting alone, and the additional gain from reference-guided selection over the generated candidates. Evaluated across all 3,406 queries with the default mixed bank, *Query-only* runs each depth model directly on the query ROI, *Outpaint* runs it on a single fixed outpainting (seed 0), and *Outpaint + Select* selects from the $N = 3$ candidates using the retrieved reference. The two blocks correspond to settings without and with query intrinsics; all configurations share identical frozen models and query ROIs. The Mean Δ row averages per-model differences from Query-only.

As shown in Tab. 1, outpainting alone improves most base models, yielding consistent gains for MoGe-3, DA3, and MoGe-2, while the other models degrade only slightly on a few metrics. This

Table 2: **Robustness to reference-bank composition.** All columns report δ_1 (%). Deeper green and red indicate larger positive and negative deviations from Query-only (gray).

	w/o query intrinsics				*w/ query intrinsics*			
Model	Query-only	Mixed	Cross-dataset	Fixed DIODE	Query-only	Mixed	Cross-dataset	Fixed DIODE
MoGe-3 (ViT-G)	63.21	82.23	80.27	80.59	80.84	84.16	83.10	83.80
Depth Anything 3	54.59	76.70	72.38	71.72	76.40	84.47	81.61	81.91
MoGe-2	62.63	82.04	80.38	80.81	83.56	86.31	85.06	85.56
UniK3D (ViT-L)	66.56	80.35	78.53	78.75	80.16	82.01	77.57	78.06
Depth Pro	63.20	73.21	72.47	72.42	76.38	77.46	78.26	78.27
UniDepth V2	69.05	78.94	76.53	76.69	78.41	80.81	75.59	76.11

matches *Sup.Mat.*'s Sec. B: generated context can occasionally favor a plausible yet incorrect scene scale. Adding the selector improves over fixed outpainting across all models in both settings. The full method produces larger average δ_1 gains without query intrinsics, but larger AbsRel reductions with intrinsics (see the Mean Δ row in Tab. 1). This difference reflects how camera calibration shapes baseline errors. Without intrinsics, severe scale ambiguity pushes many baseline predictions outside the inlier threshold; added context corrects this gross scale bias, shifting large fractions of pixels into the threshold window and boosting δ_1. With known intrinsics, baseline predictions are already better calibrated, leaving less headroom for threshold inliers; contextual cues instead refine continuous distance estimates across inliers, driving larger AbsRel reductions. Per-dataset breakdowns in *Sup.Mat.*'s Sec. D.2 confirm these trends on each of the six datasets.

While the mixed bank provides matched in-domain demonstrations, real-world deployment often encounters novel environments where matched references are unavailable. Table 2 therefore evaluates robustness to reference-bank composition across the mixed, cross-dataset, and fixed DIODE banks. Our method stays above Query-only in most model and setting combinations under both the cross-dataset and fixed DIODE banks, even when references come entirely from unrelated domains. More importantly, the mixed bank achieves the highest accuracy in 11 of the 12 model–camera settings, highlighting the benefit of diverse references with in-domain coverage. MoGe-3, DA3, and MoGe-2 improve under every bank in both settings, showing consistent receptiveness to contextual cues.

Qualitative Results. Figure 4 visualizes where generated context reshapes metric depth predictions across representative indoor and outdoor scenes. Predictions on continuous surfaces move visibly closer to ground truth, confirming that spatial context provides coherent geometric cues that correct scale drift even under known camera calibration. Additional scenes are in *Sup.Mat.*'s Sec. E, and representative failure cases in *Sup.Mat.*'s Sec. F.

Table 3: **Guided selection versus direct rescaling across reference banks.** All entries report δ_1 (%). Query-only is shown in gray; evaluated configurations use a color gradient where green and red shades reflect relative performance gains and drops against the baseline, with darker shades indicating larger changes. Pale blue and yellow mark marginal gains and drops within one percentage point, respectively.

	w/o query intrinsics							*w/ query intrinsics*						
Model	Query-only	Guided selection			Direct rescaling			Query-only	Guided selection			Direct rescaling		
		Mixed	Cross-dataset	Fixed DIODE	Mixed	Cross-dataset	Fixed DIODE		Mixed	Cross-dataset	Fixed DIODE	Mixed	Cross-dataset	Fixed DIODE
MoGe-3 (ViT-G)	63.21	**82.23**	**80.27**	**80.59**	61.53	46.03	44.19	80.84	**84.16**	**83.10**	**83.80**	83.92	70.79	71.44
Depth Anything 3	54.59	**76.70**	**72.38**	**71.72**	49.94	43.44	37.88	76.40	**84.47**	**81.61**	**81.91**	84.09	68.31	60.50
MoGe-2	62.63	**82.04**	**80.38**	**80.81**	63.47	48.70	45.63	83.56	**86.31**	**85.06**	**85.56**	84.26	64.84	65.94
UniK3D (ViT-L)	66.56	80.35	**78.53**	**78.75**	**80.67**	49.72	44.89	80.16	82.01	**77.57**	**78.06**	**85.90**	62.69	51.92
Depth Pro	63.20	**73.21**	**72.47**	**72.42**	57.19	34.08	45.77	76.38	**77.46**	**78.26**	**78.27**	75.21	45.23	56.90
UniDepth V2	69.05	**78.94**	**76.53**	**76.69**	77.70	54.90	44.32	78.41	80.81	**75.59**	**76.11**	**83.46**	60.74	47.57

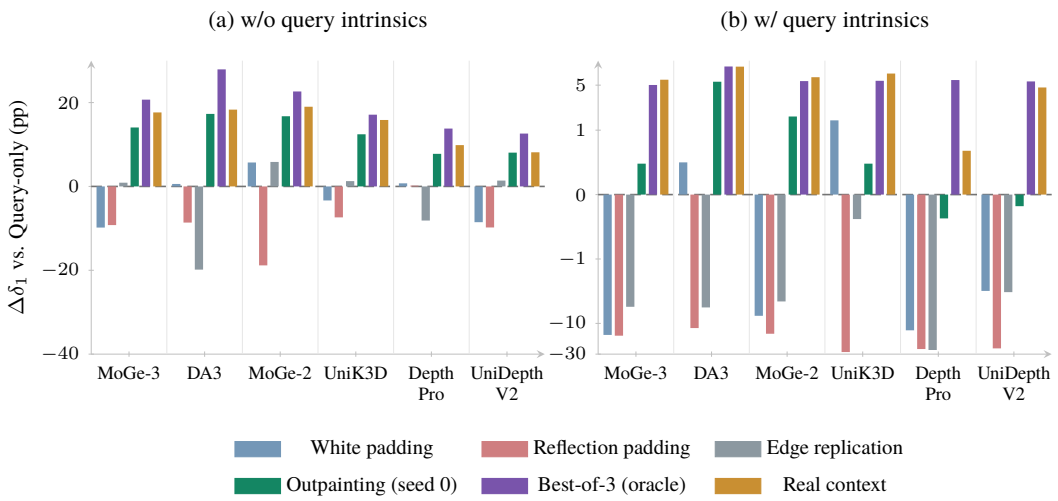

Figure 3: **Effect of context content and candidate potential.** Panel (a) uses a linear y-axis; panel (b) uses a symmetric-log y-axis (linear within ± 1 pp and logarithmic outside).

4.3 Ablation Studies

Q3 — Do the gains come from the added context or just the larger canvas?

A: From the context. Padding controls on the same canvas fall behind outpainting in 11 of 12 settings, and the Best-of-3 oracle reveals substantial untapped headroom among candidates.

Figure 3 compares outpainting with three padding controls on the same expanded canvas while preserving query pixels: white padding fills the added region with white, reflection mirrors the query across boundaries, and edge replication extends boundary pixels. Outpainting outperforms all padding controls in 11 of 12 settings, confirming that gains stem from semantic context rather than canvas expansion. Furthermore, the Best-of-3 oracle beats fixed outpainting across all models—reaching up to 10.62 δ_1 points of headroom—and matches or exceeds real context in several configurations. Multiple outpaintings thus contain superior scale candidates, motivating an effective selector to unlock these gains without ground truth (Q4).

Q4 — How should the directional bias be used: direct rescaling, or guided selection?

A: Guided selection. Direct rescaling trusts the error magnitude and collapses when the bank degrades; guided selection only uses its sign, stays close to Query-only, and keeps gains in most settings.

Table 3 compares two ways of using the same retrieved reference. **Direct rescaling** ignores the outpainting candidates and applies the reference error e_y as a scale correction, $\hat{d}_{\text{direct}} = e^{-e_y} \cdot \hat{d}_x$. **Guided selection** uses e_y only to set the preferred direction of depth change and selects a candidate without rescaling, falling back to the anchor prediction (or Query-only when query intrinsics are known; Sec. 3.2). Both share identical retrieved references and error estimates within each bank on the 2,635 non-DIODE queries. With the mixed bank, direct rescaling improves over Query-only in some configurations, but guided selection wins in most, gaining up to 22.11 δ_1 points $\uparrow$ on DA3. Under both out-of-domain banks, direct rescaling falls below Query-only in every configuration, losing up to 31.15 points $\downarrow$ on Depth Pro; guided selection keeps gains in most settings, with its worst drop at 2.82 points $\downarrow$ on UniDepth V2. Direct rescaling trusts the magnitude of e_y to transfer across scenes; guided selection uses only its sign, which is more stable under domain shift. Selector-component ablations are in *Sup.Mat.*'s Sec. D.6, and selection gains across banks in *Sup.Mat.*'s Sec. D.5.

5 Conclusion

We presented ContextDepth, which improves metric depth estimation through generated spatial context and reference-guided selection, with query pixels and pretrained weights fixed. As generative models grow stronger, we expect the headroom unlocked by their priors to widen further. The more a depth model sees, the deeper it sees.

Limitations. The field-of-view (FOV) change from outpainting can degrade models sensitive to such shifts (Secs. D.1 and D.4). Within a model's tolerance, outpainting with selection almost always improves metric depth, and selection adds value over outpainting alone. Deployment thus requires backbone-specific validation; adaptive context extent remains future work.

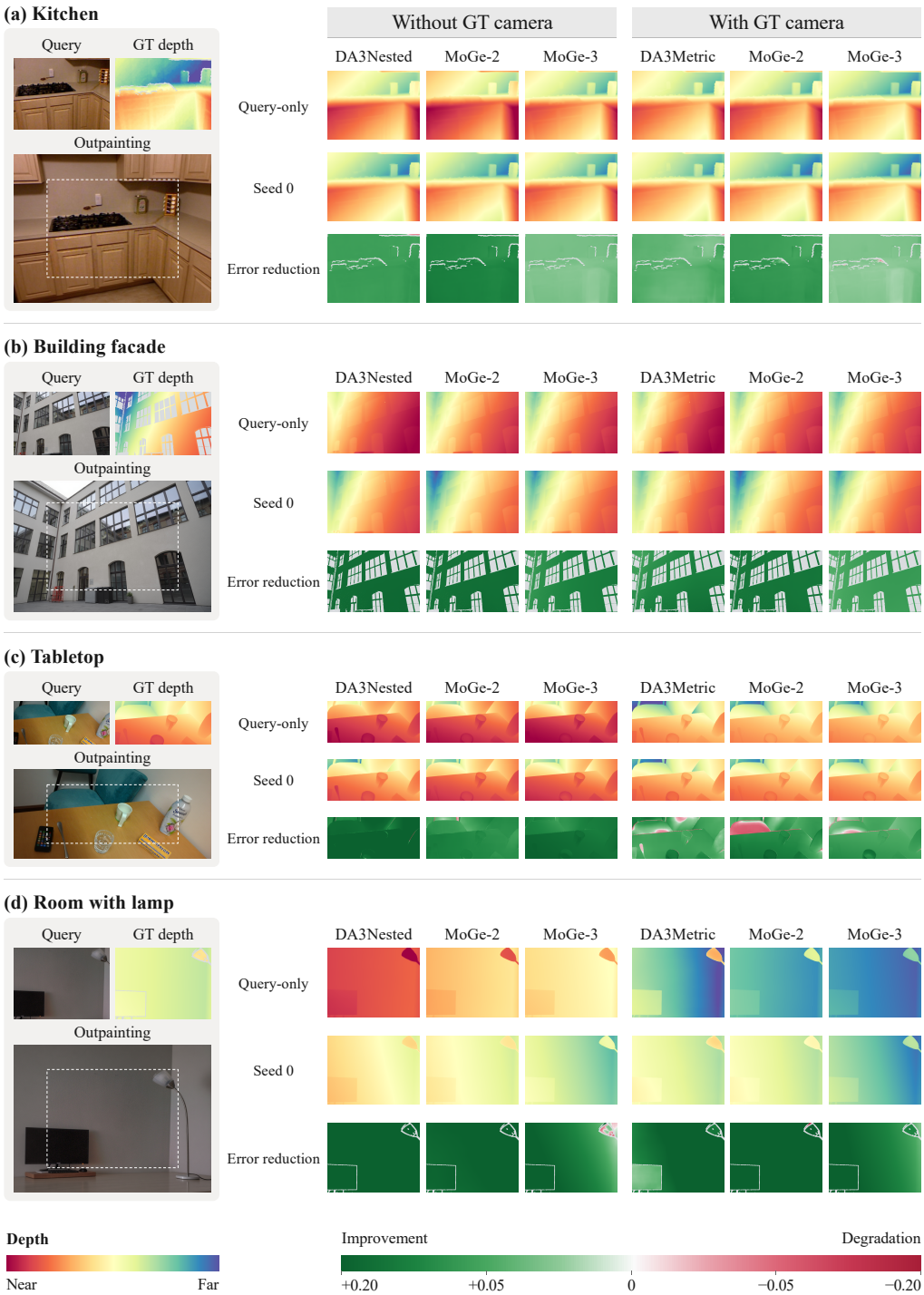

Figure 4: **Qualitative comparison of Query-only and +outpainting.** With query pixels and model weights fixed, generated context visibly reshapes predicted metric depth. In error maps, green marks error reduction and pink marks error increase for the seed-0 candidate $\hat{d}_0$ relative to Query-only, measured by $\big(|\hat{d}_x(p) - d_x(p)| - |\hat{d}_0(p) - d_x(p)|\big)/d_x(p)$, where d_x is ground-truth depth; gray marks invalid ground truth (square-root scale clipped at ± 0.20). Gains concentrate on extended surfaces such as facades, walls, and tabletops.

REFERENCES

[1] Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. Many-Shot In-Context Learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 1

[2] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual Prompting via Image Inpainting. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-Shot Transfer by Combining Relative and Metric Depth. *arXiv preprint arXiv:2302.12288*, 2023. 2, 3

[4] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. In *International Conference on Learning Representations (ICLR)*, 2025. 2, 3, 6

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 3, 13

[6] 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. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging Into Self-Supervised Monocular Depth Estimation. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[9] Yuan Gong, Xionghui Wang, Jie Wu, Shiyin Wang, Yitong Wang, and Xinglong Wu. OneReward: Unified Mask-Guided Image Generation via Multi-Task Human Preference Learning. *arXiv preprint arXiv:2508.21066*, 2025. 3, 6, 14

[10] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 2

[11] HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, Aleš Leonardis, Nassir Navab, and Benjamin Busam. On the Importance of Accurate Geometry Data for Dense 3D Vision Tasks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[12] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[13] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of CNN-Based Single-Image Depth Estimation Methods. In *European Conference on Computer Vision Workshops (ECCVw)*, 2018. 5

[14] 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. 2, 3, 6

[15] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation. arXiv preprint arXiv:1907.10326, 2021. 2

[16] Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. VisualCloze: A Universal Image Generation Framework via Visual In-Context Learning. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[17] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting Depth Anything for 4K Resolution Accurate Metric Depth Estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[18] 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 *International Conference on Learning Representations (ICLR)*, 2026. 2, 3, 6

[19] Tianci Luo, Jinpeng Wang, Shiyu Qin, Niu Lian, Yan Feng, Bin Chen, Chun Yuan, and Shu-Tao Xia. PromptHub: Enhancing Multi-Prompt Visual In-Context Learning with Locality-Aware Fusion, Concentration and Alignment. In *International Conference on Learning Representations (ICLR)*, 2026. 3

[20] Trevine Oorloff, Vishwanath Sindagi, Wele Gedara Chaminda Bandara, Ali Shafahi, Amin Ghiasi, Charan Prakash, and Reza Ardekani. Stable Diffusion Models are Secretly Good at Visual In-Context Learning. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[21] Igor Pavlovic, Thiemo Wandel, Anton Obukhov, Luca Bartolomei, Andrey Davydov, Fabio Tosi, Matteo Poggi, Sabine Süsstrunk, and Dengxin Dai. Marigold V2: Revisiting Diffusion Transformers for Monocular Depth Estimation. *Transactions on Graphics (TOG)*, 2026. 2

[22] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal Monocular Metric Depth Estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[23] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal Camera Monocular 3D Estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3, 6

[24] 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. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026. 3, 6

[25] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[26] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision (ECCV)*, 2012. 5

[27] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025. 4, 15

[28] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *arXiv preprint arXiv:1908.00463*, 2019. 5

[29] Jinpeng Wang, Tianci Luo, Yaohua Zha, Yan Feng, Ruisheng Luo, Bin Chen, Tao Dai, Long Chen, Yaowei Wang, and Shu-Tao Xia. Embracing Collaboration Over Competition: Condensing Multiple Prompts for Visual In-Context Learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[30] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[31] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate Monocular Geometry with Metric Scale and Sharp Details. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025. 2, 3, 5, 6, 14

[32] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3

[33] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards Segmenting Everything in Context. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[34] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-Context Learning Unlocked for Diffusion Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 3

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1

[36] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations (ICLR)*, 2022. 13

[37] Zhen Xu, Hongyu Zhou, Sida Peng, Haotong Lin, Haoyu Guo, Jiahao Shao, Peishan Yang, Qinglin Yang, Sheng Miao, Xingyi He, Yifan Wang, Yue Wang, Ruizhen Hu, Yiyi Liao, Xiaowei Zhou, and Hujun Bao. Towards Depth Foundation Model: Recent Trends in Vision-Based Depth Estimation. *arXiv preprint arXiv:2507.11540*, 2025. 2

[38] 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 *International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[39] Hao Yu, Haotong Lin, Jiawei Wang, Jiaxin Li, Yida Wang, Xueyang Zhang, Yue Wang, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. InfiniDepth: Arbitrary-Resolution and Fine-Grained Depth Estimation with Neural Implicit Fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2

[40] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What Makes Good Examples for Visual In-Context Learning? In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 3

Appendix

Table of Contents

A Use of Large Language Models **13**

B A Latent-Concept View of Spatial In-Context Learning **13**

C Implementation Details **14**

D Additional Experimental Results **15**

D.1 Outpainting Extent Selection with and without Query Intrinsics 15

D.2 Per-Dataset Results 16

D.3 Additional Results under AbsRel 19

D.4 Results with Full-Image Queries 20

D.5 Selector Gains across Reference Banks 21

D.6 Ablation of Selector Components 22

E Additional Qualitative Comparisons **23**

F Failure Case Analysis **23**

A Use of Large Language Models

We used large language models to assist with translation, drafting and polishing manuscript text, preparation of figures and tables, and LaTeX typesetting. The authors decide which model suggestions to adopt, make the final scientific judgments, and take responsibility for the final manuscript, experimental results, and conclusions.

B A Latent-Concept View of Spatial In-Context Learning

In language models, in-context learning (ICL) predicts query outputs from prompt demonstrations without parameter updates [5]. Spatially, the input image is the query, outpainted surroundings serve as demonstrations, and the predicted depth map is the output.

Xie et al. [36] explain ICL as implicit Bayesian inference over latent concepts: the demonstrations in a prompt are assumed to share one latent concept with the query, and the model predicts by averaging outputs over concepts weighted by their posterior probabilities. In our setting, the latent concept $z \in \mathcal{Z}$ is the global scale of the query scene. A single view fixes viewing directions but not metric distances along them, so scale information relies on learned priors. The surrounding context adds visual evidence of environment scale. Meanwhile, the query pixels constrain the relative geometry within the ROI, so context primarily shifts the scale of the prediction rather than its relative structure.

To examine context's effect on relative structure, we compare query-only and outpainted predictions after per-image scale-and-shift alignment to ground truth. As shown in Fig. 5 and Tab. 4, δ_1 changes by under 0.7 percentage points across all models and camera settings, confirming preserved relative depth within the query ROI.

Formally, let $\pi(z \mid x)$ denote the scale prior from query pixels, and let $\ell(c \mid x, z)$ score compatibility between context c and scene scale z. Bayes' rule updates this belief:

$$\pi(z \mid x, c) \;=\; \frac{\ell(c \mid x, z)\, \pi(z \mid x)}{\displaystyle\int_{\mathcal{Z}} \ell(c \mid x, z')\, \pi(z' \mid x)\, \mathrm{d}z'}. \tag{6}$$

Each candidate $\hat{x}_i$ provides one context: setting $c = \hat{x}_i$ in Eq. (6) yields posterior $\pi(z \mid x, \hat{x}_i)$. The scale estimate is a posterior average over $\mathcal{Z}$, so distinct candidates yield different depth predictions for the same query pixels under frozen weights.

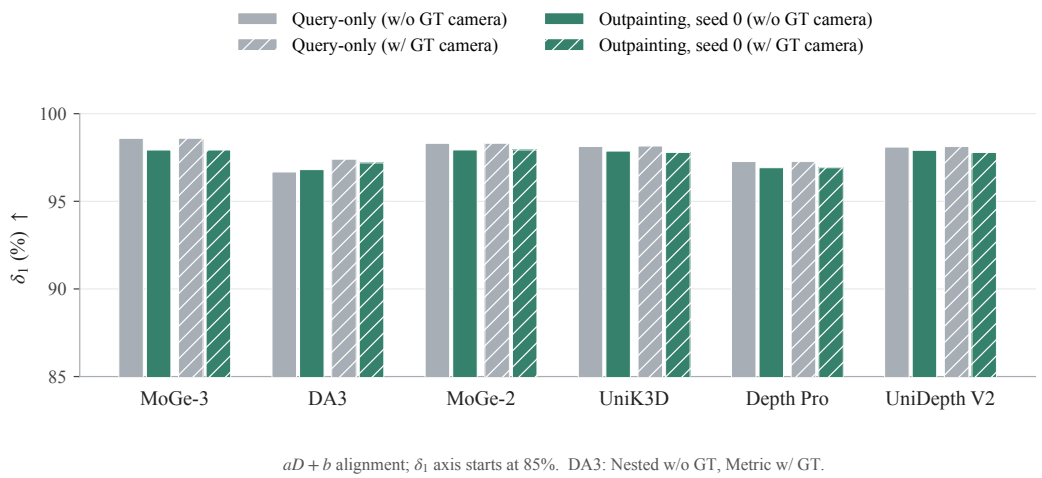

Figure 5: **Relative-depth accuracy with and without generated context.** We compare Query-only and outpainting (seed 0) after per-image scale-and-shift alignment to ground truth under both camera settings. Detailed numerical results are provided in Tab. 4.

Table 4: **Relative-depth accuracy after scale-and-shift alignment.** We report δ_1 (%) for the six models in the main comparison. Outpaint uses the fixed seed-0 candidate.

	w/o query intrinsics		w/ query intrinsics	
Model	Query-only	Outpaint	Query-only	Outpaint
MoGe-3 (ViT-G)	98.594	97.927	98.595	97.927
Depth Anything 3	96.674	96.808	97.399	97.191
MoGe-2	98.309	97.931	98.309	97.931
UniK3D (ViT-L)	98.125	97.867	98.155	97.789
Depth Pro	97.268	96.917	97.267	96.916
UniDepth V2	98.095	97.910	98.125	97.787

However, generated context may favor an incorrect scene scale, so spatial context alone is not universally beneficial. We therefore generate multiple candidate contexts and select among them using an independent reference with known depth and camera intrinsics (Sec. 3.2).

C Implementation Details

Input ROI construction. We construct input ROIs from the preprocessed evaluation images provided with MoGe-2 [31]. For each image, we retain a centered ROI spanning approximately two-thirds of its width and height, then outpaint the surrounding area. Only the cropped RGB image is provided to the outpainting model; the original surrounding pixels are not used as generation inputs. The ROI remains fixed across depth models and context variants.

OneReward outpainting settings. We generate spatial context using FLUX.1-Fill-dev with pretrained OneReward weights [9]. Inference uses bfloat16 precision and 50 denoising steps, with guidance scale set to 1.0 and true cfg set to 4.0. VAE tiling and slicing are enabled. All inputs use the positive prompt "high-definition, perfect composition" and the negative prompt "nsfw". We generate three candidates per input image, with random seeds determined by the image identifier and candidate index. The same outpaintings are reused across depth models and both camera settings. Outpainting is performed on an NVIDIA A100-SXM4-80GB GPU.

The input ROI is centered in the outpainting canvas. The added regions are initialized to white, and the generation mask is set to 0 inside the ROI and 255 outside it. After generation, we copy the pre-outpainting ROI pixels back to their corresponding locations and verify pixelwise equality, ensuring that outpainting changes only the context outside the ROI.

Pretrained depth models and camera settings. We use the official implementations and pretrained weights of six depth models through their native inference interfaces, keeping their network

architectures and parameters fixed. For Depth Anything 3, we use DA3Nested-Giant-Large-1.1 when input intrinsics are unknown and DA3Metric-Large when they are known. The remaining models use the same weights in both settings: MoGe-2 (ViT-L), MoGe-3 (ViT-G), UniDepth V2 (ViT-L/14), UniK3D (ViT-L), and the official pretrained Depth Pro model.

When input intrinsics are unknown, each model follows its official inference procedure without camera inputs. When input intrinsics are known, we use the camera parameters in the form supported by each model's official implementation. Camera parameters are mapped to the appropriate coordinate system according to input resizing, ROI placement, and model preprocessing.

Selector implementation details. Reference retrieval uses a frozen SigLIP 2 SO400M image encoder [27], with a patch size of 16 and an input resolution of 384×384. Images undergo the encoder's standard preprocessing, followed by attention pooling to obtain 1152-dimensional descriptors. We L2-normalize these descriptors and compute cosine similarity.

Reference depth predictions follow the corresponding model's known-intrinsics configuration, and reference errors are computed using Eq. (1). For DA3Nested, we retain the same Nested checkpoint and perform metric calibration using the known reference focal length. Camera distances are computed from the logarithms of focal lengths normalized by the corresponding ROI dimensions. Depth Pro uses only the horizontal focal component, whereas the other models use both horizontal and vertical components. For UniK3D, which predicts general camera rays, we fit an equivalent pinhole camera by least squares within the visible ROI and use its focal lengths for the Selector's camera comparison. This fitting uses no input ground-truth depth and is not fed back into depth prediction in the unknown-intrinsics setting.

Candidate depth changes are computed on common ROI pixels where all three candidate predictions are finite and greater than 10^{-6}.

D Additional Experimental Results

D.1 Outpainting Extent Selection with and without Query Intrinsics

Q5 — How do we choose the outpainting extent across model–camera settings?
A: Use 25% by default and 10% for models sensitive to expansion. We calibrate the extent on ETH3D without query intrinsics; the chosen configurations retain gains on the other five datasets.

The main paper uses a 25% outpainting extent per ROI side by default. The query spans about two-thirds of the original width and height, so this extent restores roughly the original image-to-ROI proportions. With query intrinsics, camera parameters are given and need not be inferred from content, so all models keep this default.

Without query intrinsics, a changed canvas can distort the model's estimate of camera parameters and scene scale. We therefore calibrate the extent on ETH3D, which has the widest field of view among the six datasets. The minimum extent considered is 10% per side, small enough to retain some context. We compare per-side extents of 10%, 15%, 20%, and 25% by extracting nested windows around the fixed ROI from the same seed-0 outpainting. Each window goes through the frozen depth model alone, without regenerating context or selecting candidates. Metrics are computed within the same ROI and averaged over the 454 ETH3D queries (Tab. 5).

From 15% to 25%, MoGe-3 and MoGe-2 vary little, while DA3 keeps improving. UniK3D, Depth Pro, and UniDepth V2 show the opposite trend: AbsRel rises progressively, and δ_1 at 25% falls below its value at 10%. We therefore keep 25% for MoGe-3, DA3, and MoGe-2, and use a conservative 10% for the other three.

We then fix these extents and repeat the mixed-bank protocol of Tab. 1, excluding ETH3D, which was used for calibration. Table 6 covers the remaining 2,952 queries from the five other datasets. The full method beats Query-only on both metrics in every model–camera setting, and selection further improves both metrics over fixed outpainting (seed 0). The aggregate gains of Tab. 1 therefore do not hinge on ETH3D.

Table 5: **Outpainting extent on ETH3D without query intrinsics.** Column headings indicate the percentage added to each side of the query ROI. **Bold** and underlined values mark the best and second-best results among the displayed extents for each model and metric; δ_1 is reported in percent.

Model	δ_1 (%)↑				AbsRel↓			
	10%	15%	20%	25%	10%	15%	20%	25%
MoGe-3 (ViT-G)	80.60	81.70	**81.84**	81.74	0.1479	**0.1453**	0.1465	0.1487
Depth Anything 3	62.23	67.78	74.83	**80.04**	0.1936	0.1898	0.1778	**0.1721**
MoGe-2	85.23	86.48	**87.19**	86.83	0.1137	**0.1136**	0.1139	0.1182
UniK3D (ViT-L)	**79.97**	77.01	73.70	71.07	**0.1646**	0.1751	0.1895	0.2079
Depth Pro	42.48	**43.63**	38.94	36.89	**0.3293**	0.3342	0.3568	0.3657
UniDepth V2	**72.13**	68.65	66.26	62.26	**0.1887**	0.2106	0.2268	0.2447

Table 6: **Metric depth estimation excluding ETH3D.** δ_1 is reported in percent. Green, yellow-green, and red denote the best, second-best, and worst results per model, camera setting, and metric; shading reflects absolute metric quality within each setting and metric. **Bold** and underlined entries mark the best and second-best results. Mean Δ averages changes from Query-only over six models.

Model	*w/o query intrinsics*						*w/ query intrinsics*					
	Query-only		Outpaint		Outpaint + Select		Query-only		Outpaint		Outpaint + Select	
	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓
MoGe-3 (ViT-G)	60.80	0.2410	75.96	0.2365	**78.45**	**0.2125**	76.81	0.2506	77.62	0.2426	**80.22**	**0.2069**
Depth Anything 3	50.32	0.2392	66.69	0.2086	**72.97**	**0.1796**	73.03	0.2157	79.11	0.1777	**80.68**	**0.1597**
MoGe-2	57.01	0.2204	75.55	0.2091	**77.13**	**0.1903**	79.17	0.2079	80.85	0.1976	**82.17**	**0.1689**
UniK3D (ViT-L)	59.46	0.2830	74.51	0.2813	**76.21**	**0.2414**	74.24	0.2653	75.46	0.2489	**77.77**	**0.1920**
Depth Pro	57.86	0.2423	66.61	0.2358	**68.20**	**0.2199**	70.83	0.2165	71.86	0.2084	**73.74**	**0.1857**
UniDepth V2	59.59	0.3124	69.82	0.2995	**71.16**	**0.2727**	71.01	0.2981	71.65	0.2948	**73.96**	**0.2333**
Mean Δ	–	–	**+14.02**	**−0.0112**	**+16.51**	**−0.0370**	–	–	**+1.91**	**−0.0140**	**+3.91**	**−0.0513**

D.2 Per-Dataset Results

Q6 — Do the aggregate gains in Tab. 1 also hold at the dataset level?
A: Yes, outpainting helps and selection improves it further. Outpainting with seed 0 outperforms Query-only in most configurations, and reference-guided selection further improves most results. These gains span datasets, models, and camera settings.

Table 7 breaks down the comparison in Tab. 1 by dataset, following the same evaluation protocol. Across six datasets, six models, and two camera settings, the full method ranks first or second among the three methods in 130 of the 144 entries across both metrics (90.3%). This indicates competitive performance across diverse datasets and camera settings.

Relative to Query-only, the largest δ_1 decrease across these configurations is 6.66 percentage points ↓, whereas the largest increase reaches 66.68 percentage points ↑; for AbsRel, the largest increase is 0.0943 ↑ and the largest reduction is 0.1889 ↓. For both metrics, the maximum improvement exceeds twice the maximum degradation in magnitude, demonstrating the potential to correct large metric-depth errors. Together with the largely preserved relative geometry observed in Sec. B, these gains are consistent with more accurate estimation of absolute scene scale.

Compared with fixed outpainting (seed 0), reference-guided selection increases δ_1 in 55 of the 72 configurations and reduces AbsRel in 58, while improving both metrics in the aggregate results for all twelve model–camera settings. This shows that reference guidance can exploit quality differences among outpainting candidates to further improve metric depth predictions.

Table 7: **Per-dataset metric depth results without query intrinsics.** Green, yellow-green, and red denote the best, second-best, and worst results within each model, dataset, and metric; shading reflects metric quality. **Bold** and underlined values mark the best and second-best results, ranked before rounding. n is the query count; All denotes the average over all queries.

(a) Threshold accuracy $\delta_1\uparrow$ (%)

Model	Method	DIODE $n = 771$	ETH3D $n = 454$	HAMMER $n = 775$	KITTI $n = 652$	NYUv2 $n = 654$	iBims-1 $n = 100$	All $n = 3{,}406$
MoGe-3 (ViT-G)	Query-only	60.83	74.87	43.33	45.67	**94.30**	75.48	62.67
	Outpaint	66.45	81.74	62.99	86.51	90.67	84.72	76.73
	Outpaint + Select	**67.92**	**82.49**	**66.20**	**88.70**	93.38	**90.12**	**78.99**
Depth Anything 3	Query-only	39.49	56.66	60.71	7.42	90.14	72.76	51.17
	Outpaint	61.45	**80.04**	62.39	52.35	89.43	85.24	68.47
	Outpaint + Select	**63.99**	79.35	**69.36**	**65.65**	**92.64**	**89.30**	**73.82**
MoGe-2	Query-only	52.40	81.85	44.86	36.48	**95.03**	71.87	60.32
	Outpaint	64.37	**86.83**	64.48	84.77	91.53	83.09	77.06
	Outpaint + Select	**65.42**	85.75	**65.34**	**87.65**	92.97	**86.56**	**78.28**
UniK3D (ViT-L)	Query-only	49.95	**84.53**	**62.57**	27.90	**93.80**	89.84	62.80
	Outpaint	61.02	79.97	55.53	93.84	91.12	91.10	75.24
	Outpaint + Select	**63.04**	77.87	58.44	**94.58**	92.11	**91.55**	**76.43**
Depth Pro	Query-only	29.70	41.01	61.16	61.08	81.70	72.55	55.62
	Outpaint	34.71	42.48	58.79	85.25	92.06	85.06	63.39
	Outpaint + Select	**37.31**	**44.86**	**61.39**	**85.56**	**92.55**	**86.59**	**65.08**
UniDepth V2	Query-only	38.20	**78.16**	**56.02**	55.39	89.30	85.38	62.07
	Outpaint	45.93	72.13	51.92	94.35	91.78	**89.31**	70.13
	Outpaint + Select	**46.40**	74.30	54.84	**94.98**	**93.33**	88.22	**71.58**

(b) Absolute relative error (AbsRel$\downarrow$)

Model	Method	DIODE $n = 771$	ETH3D $n = 454$	HAMMER $n = 775$	KITTI $n = 652$	NYUv2 $n = 654$	iBims-1 $n = 100$	All $n = 3{,}406$
MoGe-3 (ViT-G)	Query-only	0.2215	0.1652	**0.4262**	0.2159	**0.0843**	0.1433	0.2309
	Outpaint	0.2061	0.1487	0.4991	0.1126	0.1009	0.1308	0.2248
	Outpaint + Select	**0.1966**	**0.1474**	0.4350	**0.1053**	0.0895	**0.1135**	**0.2038**
Depth Anything 3	Query-only	0.2930	0.2157	**0.2156**	0.3583	0.0994	0.1457	0.2361
	Outpaint	0.2175	**0.1721**	0.2958	0.2101	0.1058	0.1270	0.2038
	Outpaint + Select	**0.2113**	0.1735	0.2383	**0.1709**	**0.0917**	**0.1107**	**0.1788**
MoGe-2	Query-only	0.2250	0.1237	**0.3275**	0.2347	**0.0820**	0.1656	0.2075
	Outpaint	**0.1948**	**0.1182**	0.4024	0.1158	0.1014	0.1350	0.1970
	Outpaint + Select	0.1948	0.1236	0.3466	**0.1092**	0.0914	**0.1189**	**0.1814**
UniK3D (ViT-L)	Query-only	0.4921	**0.1293**	**0.3087**	0.2300	**0.0862**	0.1042	0.2625
	Outpaint	0.4380	0.1646	0.4463	0.1050	0.1044	**0.1019**	0.2658
	Outpaint + Select	**0.3380**	0.1673	0.4030	**0.0992**	0.0989	0.1040	**0.2315**
Depth Pro	Query-only	0.3575	0.3198	**0.2788**	0.1845	0.1310	0.1771	0.2527
	Outpaint	0.3338	0.3293	0.3643	0.1185	0.0998	0.1393	0.2483
	Outpaint + Select	**0.3198**	**0.3186**	0.3228	**0.1164**	**0.0975**	**0.1275**	**0.2331**
UniDepth V2	Query-only	0.5536	**0.1572**	**0.3555**	0.1953	0.1221	0.1259	0.2917
	Outpaint	0.4841	0.1887	0.4769	0.0934	0.1062	0.1099	0.2847
	Outpaint + Select	**0.4519**	0.1833	0.4205	**0.0874**	**0.0961**	**0.1085**	**0.2608**

Table 7: **Per-dataset metric depth results with query intrinsics (continued).**
Colors and notation follow the preceding page.

(a) Threshold accuracy $\delta_1\uparrow$ (%)

Model	Method	DIODE $n = 771$	ETH3D $n = 454$	HAMMER $n = 775$	KITTI $n = 652$	NYUv2 $n = 654$	iBims-1 $n = 100$	All $n = 3{,}406$
MoGe-3 (ViT-G)	Query-only	69.87	**88.46**	52.73	91.76	**96.55**	90.24	78.36
	Outpaint	69.88	86.78	56.45	93.07	94.38	**90.96**	78.84
	Outpaint + Select	**71.15**	87.68	**64.02**	**93.31**	95.50	90.36	**81.21**
Depth Anything 3	Query-only	65.77	80.20	42.66	91.83	**96.35**	89.45	73.99
	Outpaint	**71.05**	83.11	60.60	92.83	94.85	**92.42**	79.64
	Outpaint + Select	69.39	**83.47**	**67.09**	**93.04**	96.23	90.82	**81.05**
MoGe-2	Query-only	69.59	88.39	64.71	89.01	**96.33**	88.56	80.39
	Outpaint	70.11	**89.60**	68.90	92.82	94.13	**91.52**	82.02
	Outpaint + Select	**70.96**	87.15	**72.51**	**93.11**	94.98	88.32	**82.83**
UniK3D (ViT-L)	Query-only	62.93	**89.39**	52.99	90.90	**93.24**	**93.27**	76.26
	Outpaint	66.82	85.04	50.64	**95.93**	92.79	87.77	76.74
	Outpaint + Select	**67.05**	84.13	**60.21**	95.34	92.28	87.10	**78.62**
Depth Pro	Query-only	39.89	**50.49**	69.21	**87.53**	90.09	**86.97**	68.12
	Outpaint	44.01	41.00	68.74	87.50	90.55	86.60	67.75
	Outpaint + Select	**46.98**	49.87	**72.97**	87.42	**90.61**	86.66	**70.56**
UniDepth V2	Query-only	54.03	**85.10**	47.19	95.22	92.15	**90.49**	72.89
	Outpaint	**57.33**	79.61	46.73	**95.61**	91.49	89.30	72.71
	Outpaint + Select	56.55	84.13	**55.14**	94.95	**93.81**	87.39	**75.32**

(b) Absolute relative error (AbsRel$\downarrow$)

Model	Method	DIODE $n = 771$	ETH3D $n = 454$	HAMMER $n = 775$	KITTI $n = 652$	NYUv2 $n = 654$	iBims-1 $n = 100$	All $n = 3{,}406$
MoGe-3 (ViT-G)	Query-only	0.2335	0.1468	0.5735	0.0925	**0.0690**	**0.0985**	0.2368
	Outpaint	0.2167	**0.1300**	0.5507	0.0874	0.0844	0.1019	0.2276
	Outpaint + Select	**0.2066**	0.1323	**0.4330**	**0.0844**	0.0778	0.1004	**0.1970**
Depth Anything 3	Query-only	0.2115	0.1398	0.4573	0.0966	**0.0725**	**0.0898**	0.2056
	Outpaint	**0.1903**	0.1382	0.3340	0.0902	0.0776	0.0946	0.1724
	Outpaint + Select	0.1929	**0.1287**	**0.2684**	**0.0873**	0.0738	0.0938	**0.1555**
MoGe-2	Query-only	0.2091	0.1142	0.4192	0.1105	**0.0683**	0.1095	0.1954
	Outpaint	0.2018	**0.1056**	0.3863	0.0991	0.0824	**0.1003**	0.1854
	Outpaint + Select	**0.2015**	0.1144	**0.2849**	**0.0941**	0.0776	0.1033	**0.1616**
UniK3D (ViT-L)	Query-only	0.4357	**0.1186**	0.3882	0.1274	**0.0829**	**0.0903**	0.2457
	Outpaint	0.3455	0.1445	0.4450	**0.0730**	0.0963	0.1279	0.2349
	Outpaint + Select	**0.2830**	0.1390	**0.2865**	0.0779	0.0958	0.1316	**0.1850**
Depth Pro	Query-only	0.3277	0.2784	0.2941	0.1172	0.1059	**0.1294**	0.2248
	Outpaint	0.3128	0.3224	0.2791	0.1161	**0.1045**	0.1356	0.2236
	Outpaint + Select	**0.2945**	**0.2784**	**0.2102**	**0.1157**	0.1060	0.1349	**0.1981**
UniDepth V2	Query-only	0.5305	0.1455	0.4262	0.0931	0.1077	**0.0946**	0.2777
	Outpaint	0.4615	0.1718	0.4981	**0.0746**	0.1045	0.1151	0.2784
	Outpaint + Select	**0.3983**	**0.1438**	**0.3335**	0.0805	**0.0890**	0.1242	**0.2214**

D.3 Additional Results under AbsRel

Due to space constraints, the main paper reports only δ_1 for the reference-bank comparison (Tab. 2) and the two ablation studies (Fig. 3 and Tab. 3). Here, we follow the corresponding experimental settings and additionally report AbsRel to examine whether these conclusions also hold under a continuous relative-error metric.

Q7 — Do the conclusions drawn from δ_1 in the main paper also hold for AbsRel?

A: Yes, the main conclusions still hold. AbsRel also supports the advantage of the mixed reference bank and the effectiveness of cross-dataset references (Tab. 2), the benefits of generated context and the headroom among candidates (Fig. 3), and the greater robustness of guided selection over direct rescaling under restricted reference banks (Tab. 3).

D.3.1 Reference-Bank Composition

Table 8 follows the experimental settings of Tab. 2 and reports AbsRel for the three reference banks.

The mixed bank achieves the lowest AbsRel among the three banks in almost all model–camera settings. This agrees with the conclusions drawn from δ_1 in Tab. 2, further supporting the interpretation that in-domain references provide more relevant guidance.

Both the cross-dataset and fixed DIODE banks achieve lower AbsRel than Query-only in most settings. This further supports the conclusion in Tab. 2: in-domain references are generally more beneficial, but are not necessary for gains. Our method still improves metric depth estimation in most settings without same-dataset references.

Table 8: **Reference-bank composition evaluated with AbsRel.** Gray denotes Query-only; green and red mark lower and higher AbsRel relative to Query-only, respectively, with darker shades indicating larger changes. Pale blue and yellow mark marginal reductions and increases of at most 0.01. **Bold** marks the lowest AbsRel within each model and camera setting.

AbsRel↓	*w/o query intrinsics*				*w/ query intrinsics*			
Model	Query-only	Mixed	Cross-dataset	Fixed DIODE	Query-only	Mixed	Cross-dataset	Fixed DIODE
MoGe-3 (ViT-G)	0.2336	**0.2059**	0.2136	0.2116	0.2377	**0.1941**	0.1994	0.1957
Depth Anything 3	0.2194	**0.1692**	0.1848	0.1869	0.2039	**0.1446**	0.1563	0.1570
MoGe-2	0.2024	**0.1775**	0.1843	0.1825	0.1914	**0.1499**	0.1577	0.1572
UniK3D (ViT-L)	**0.1953**	0.2004	0.2131	0.2115	0.1901	**0.1563**	0.1925	0.1898
Depth Pro	0.2220	**0.2077**	0.2149	0.2155	0.1947	0.1698	0.1707	**0.1687**
UniDepth V2	0.2150	**0.2049**	0.2216	0.2211	0.2038	**0.1696**	0.2118	0.2124

D.3.2 Context Content and Candidate Potential

Figure 6 follows the experimental settings of Fig. 3 and reports AbsRel for the different context constructions and the Best-of-3 oracle.

Outpainting generally achieves lower AbsRel than white padding, reflection padding, and edge replication. This agrees with the δ_1 results in the main paper, further supporting the conclusion that the gains come from added context content rather than canvas expansion alone.

Compared with fixed outpainting (seed 0), the Best-of-3 oracle further reduces AbsRel. This shows that the potential gains from candidate selection observed in the main paper also extend to continuous relative error, further motivating an effective selection mechanism for multiple outpaintings.

D.3.3 Reference-Guided Selection versus Direct Rescaling

Table 9 follows the experimental settings of Tab. 3 and reports AbsRel for guided selection and direct rescaling across the three reference banks.

Under the cross-dataset and fixed DIODE banks, guided selection generally achieves lower AbsRel than direct rescaling and retains improvements over Query-only in most settings. This agrees with the conclusions drawn from δ_1 in Tab. 3, further supporting the greater robustness of guided selection under restricted reference banks.

Across these configurations, the largest AbsRel increase over Query-only is much smaller for guided selection than for direct rescaling. This further supports the design rationale in Tab. 3: selecting candidates using only the direction of the reference error can mitigate the adverse effects of reference-domain shifts, compared with directly transferring its magnitude as a scale correction.

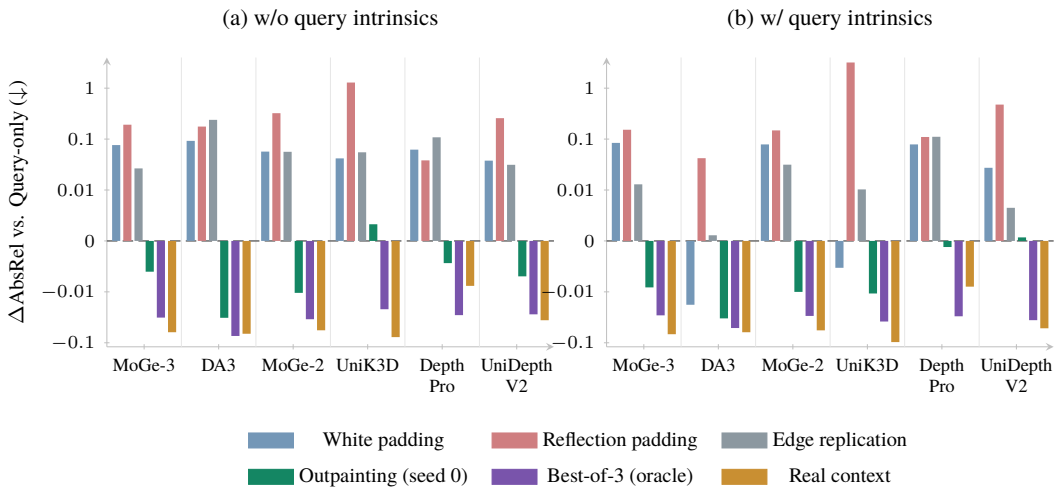

Figure 6: **Effect of context content and candidate potential on AbsRel.** Bars show AbsRel changes relative to Query-only; negative values indicate improvement. Blue, red, and gray denote white padding, reflection padding, and edge replication; green, purple, and orange denote fixed outpainting (seed 0), the Best-of-3 oracle, and real context. Panels (a)/(b) are without/with query intrinsics; both use symmetric-log y-axes.

Table 9: **Guided selection versus direct rescaling under AbsRel.** All entries report AbsRel↓. Gray denotes Query-only; green and red indicate decreases and increases relative to Query-only, with darker shades indicating larger changes. Pale blue and yellow mark decreases and increases of at most 0.01, respectively. **Bold** and underlining mark the lower and higher AbsRel between the two strategies within each model, reference bank, and camera setting.

Model	Query-only	Guided selection			Direct rescaling		
		Mixed	Cross-dataset	Fixed DIODE	Mixed	Cross-dataset	Fixed DIODE
(a) w/o query intrinsics							
MoGe-3 (ViT-G)	0.2336	0.2059	**0.2136**	**0.2116**	**0.2031**	0.2734	0.2910
Depth Anything 3	0.2194	**0.1692**	**0.1848**	**0.1869**	0.2309	0.2509	0.2896
MoGe-2	0.2024	**0.1775**	**0.1843**	**0.1825**	0.1883	0.2506	0.2588
UniK3D (ViT-L)	0.1953	0.2004	**0.2131**	**0.2115**	**0.1554**	0.2652	0.2873
Depth Pro	0.2220	**0.2077**	**0.2149**	**0.2155**	0.2216	0.3646	0.3035
UniDepth V2	0.2150	0.2049	**0.2216**	**0.2211**	**0.1623**	0.2865	0.3252
(b) w/ query intrinsics							
MoGe-3 (ViT-G)	0.2377	0.1941	**0.1994**	**0.1957**	**0.1578**	0.2089	0.2151
Depth Anything 3	0.2039	0.1446	**0.1563**	**0.1570**	**0.1378**	0.1951	0.2159
MoGe-2	0.1914	**0.1499**	**0.1577**	**0.1572**	0.1519	0.2134	0.2130
UniK3D (ViT-L)	0.1901	0.1563	**0.1925**	**0.1898**	**0.1344**	0.2607	0.2857
Depth Pro	0.1947	**0.1698**	**0.1707**	**0.1687**	0.1767	0.4112	0.2875
UniDepth V2	0.2038	0.1696	**0.2118**	**0.2124**	**0.1385**	0.2867	0.3237

D.4 Results with Full-Image Queries

Q8 — Can full-image queries still benefit from context expansion?
A: Yes, but gains depend on the model and query intrinsics. With query intrinsics, both δ_1 and AbsRel improve for all models. Without intrinsics, a few models sensitive to canvas expansion show degradation, consistent with the observations in Sec. D.1, the best expansion extend varies (10%-25%) by model.

We follow the comparison protocol and mixed reference bank of Tab. 1, but use complete images as queries, without center cropping. Every model uses a 10% outpainting extent per side of the original image. Predictions are evaluated on valid pixels within the original image region. This tests whether context expansion still helps when the query retains all original content.

With query intrinsics, the full method (Outpaint + Select) beats Query-only on both metrics for all models (Tab. 10). Without intrinsics, MoGe-3, DA3, and MoGe-2 also improve on both metrics.

Table 10: **Metric depth estimation with full-image queries.** δ_1 is reported in percent. Green, yellow-green, and red denote the best, second-best, and worst results within each model, camera setting, and metric; shading reflects absolute metric quality within each camera setting and metric. **Bold** and underlined entries mark the best and second-best results.

	Query-only		Outpaint (seed 0)		Outpaint + Select	
Model	δ_1 (%) ↑	AbsRel ↓	δ_1 (%) ↑	AbsRel ↓	δ_1 (%) ↑	AbsRel ↓
Without query intrinsics						
MoGe-3 (ViT-G)	79.63	0.1638	84.77	0.1661	**85.40**	**0.1599**
DA3	73.06	0.1619	75.69	0.1612	**78.67**	**0.1491**
MoGe-2	78.35	0.1570	84.07	0.1503	**84.64**	**0.1458**
UniK3D (ViT-L)	**80.86**	**0.1621**	71.38	0.2314	72.52	0.2217
Depth Pro	**62.09**	**0.2803**	48.48	0.3410	49.85	0.3313
UniDepth V2	**69.48**	**0.2607**	65.52	0.2916	66.32	0.2801
With query intrinsics						
MoGe-3 (ViT-G)	84.83	0.1630	85.81	0.1681	**86.53**	**0.1528**
DA3	84.17	0.1337	85.21	0.1218	**86.65**	**0.1181**
MoGe-2	86.77	0.1441	87.36	0.1408	**87.57**	**0.1335**
UniK3D (ViT-L)	83.94	0.1478	82.73	0.1571	**84.87**	**0.1454**
Depth Pro	67.69	0.2197	66.59	0.2159	**68.97**	**0.2045**
UniDepth V2	77.26	0.2269	76.40	0.2224	**79.29**	**0.1971**

Gains from generated context thus go beyond merely restoring what cropping removed. Even complete images benefit.

Without query intrinsics, UniK3D, Depth Pro, and UniDepth V2 instead degrade on both metrics. This matches their sensitivity to larger outpainting extents on ETH3D (Sec. D.1). The likely cause is a distribution shift at test time. A complete image already covers a wider field of view than a cropped ROI, and outpainting extends the canvas further, beyond the original image boundaries. Such wide views deviate from the camera and scale priors these frozen models learned during training. Without intrinsics, the models must estimate camera parameters from the image itself, so the shift disrupts this estimation and corrupts absolute scale, offsetting the benefit of added context. This split tracks architecture. Depth Pro regresses FOV from the image, UniDepth V2 predicts a dense camera representation that prompts its depth decoder, and UniK3D predicts a pencil of rays that unprojects depth into 3D. Metric scale in the three degrading models therefore passes through an explicit, input-adaptive camera estimate—the very quantity the FOV shift distorts. The other models lack this intermediate: MoGe-3 and MoGe-2 regress point maps under a fixed virtual camera, and DA3 conditions depth on a shared learnable camera token when intrinsics are unknown. Even for these degrading models, reference-guided selection still increases δ_1 and reduces AbsRel relative to fixed outpainting (seed 0); it does so in every model–camera setting.

These results mark the boundary of our method. Outpainting is not a universal cure: the sizable field-of-view change it introduces can degrade models sensitive to such shifts. What we can claim is narrower: when the field-of-view change stays within a model's tolerance, outpainting with selection almost always improves metric depth, and selection consistently adds value over outpainting alone.

D.5 Selector Gains across Reference Banks

Q9 — How does reference-bank composition affect the gains from selection?

A: Cross-dataset references remain useful; in-domain references yield more consistent gains. Compared with fixed outpainting (seed 0), the mixed bank improves both δ_1 and AbsRel in all model–camera settings. The cross-dataset bank still reduces AbsRel in all settings, while mean δ_1 remains close to that of seed 0 (Tab. 11).

Table 11 follows the reference-bank comparison protocol in Tab. 2 and reports the additional gains from selection with the mixed and cross-dataset banks relative to fixed outpainting (seed 0).

With the mixed bank, selection increases δ_1 and reduces AbsRel in all model–camera settings. This shows that reference guidance can exploit quality differences among outpainting candidates to further improve depth estimation beyond a single fixed outpainting.

With the cross-dataset bank, AbsRel also improves in all settings, while δ_1 improves or decreases by less than one percentage point in most settings; mean δ_1 remains close to that of seed 0. These results show that reductions in relative error do not depend entirely on in-domain references, while their availability helps yield more consistent gains on both metrics, further supporting reference-guided candidate selection.

Table 11: **Selector gains over fixed outpainting (seed 0).** Entries report selected scores minus seed-0 scores; $\Delta\delta_1$ is measured in percentage points (pp). Green and red denote gains and losses, with darker shades indicating larger changes. Pale blue and yellow mark gains and losses of at most 1 pp in δ_1 or 0.01 in AbsRel, respectively. **Mean** averages the twelve model–camera settings with equal weights.

Model	Mixed		Cross-dataset	
	$\Delta\delta_1$ (pp)↑	ΔAbsRel↓	$\Delta\delta_1$ (pp)↑	ΔAbsRel↓
w/o query intrinsics				
MoGe-3 (ViT-G)	+2.49	−0.0244	+0.53	−0.0167
DA3Nested	+6.17	−0.0305	+1.85	−0.0149
MoGe-2	+1.27	−0.0202	−0.39	−0.0133
UniK3D (ViT-L)	+0.95	−0.0150	−0.87	−0.0023
Depth Pro	+1.43	−0.0156	+0.69	−0.0084
UniDepth V2	+1.73	−0.0216	−0.69	−0.0048
w/ query intrinsics				
MoGe-3 (ViT-G)	+2.70	−0.0367	+1.64	−0.0314
DA3Metric	+2.31	−0.0226	−0.55	−0.0109
MoGe-2	+0.80	−0.0306	−0.44	−0.0229
UniK3D (ViT-L)	+2.36	−0.0463	−2.07	−0.0101
Depth Pro	+2.77	−0.0276	+3.57	−0.0268
UniDepth V2	+3.59	−0.0553	−1.63	−0.0130
Mean	**+2.38**	**−0.0289**	**+0.14**	**−0.0146**

D.6 Ablation of Selector Components

Q10 — Do camera matching and direction filtering each improve candidate selection?

A: Both contribute, and reference-error signs provide useful guidance. With the candidate pool fixed, camera matching improves over random selection. Adding direction filtering further reduces AbsRel for all models and increases δ_1 for most. Holding the other selection conditions fixed, actual reference-error signs likewise reduce AbsRel for all models and increase δ_1 for most compared with shuffled signs (Tab. 12).

Following the setting without query intrinsics and with the mixed reference bank in Tab. 1, we compare random selection, camera matching alone, and the full selector using the same candidate pool. We additionally use a sign-shuffling control to examine the role of reference-error signs (Tab. 12).

Camera matching alone improves both δ_1 and AbsRel over random selection for all models, indicating that reference camera information provides useful cues for candidate ranking. Adding direction filtering further reduces AbsRel for all models and increases δ_1 for most, showing that reference error directions provide additional information beyond camera matching.

With the other selection conditions fixed, actual reference-error signs yield lower AbsRel for all models and higher δ_1 for most than signs shuffled within each dataset. Since this control preserves the proportions of positive and negative signs and uses the same filtering rule, the results further support the value of the correspondence between reference-error signs and individual samples for candidate selection.

Table 12: **Selector component ablation without query intrinsics.** All settings use the mixed bank and the same three-candidate pool for 3,406 queries. Camera uses camera matching; Full adds reference-error direction filtering. Actual−shuffled reports actual-sign scores minus their exact expectation under within-dataset sign permutations, preserving sign proportions and all other selection inputs. **Bold**/underline mark the best/second-best of the three strategies. Gray marks Random. Green/red indicate gains/losses versus Random (strategies) or shuffled signs (sign control); darker shades indicate larger changes. Pale blue/yellow mark gains/losses of at most 1 pp in δ_1 or 0.01 in AbsRel.

Model	Random		Camera		Full		Actual − shuffled	
	δ_1 (%)↑	AbsRel↓	δ_1 (%)↑	AbsRel↓	δ_1 (%)↑	AbsRel↓	$\Delta\delta_1$ (pp)↑	ΔAbsRel↓
MoGe-3 (ViT-G)	77.25	0.2205	**79.01**	0.2179	78.99	**0.2038**	+0.479	−0.00295
Depth Anything 3	69.08	0.1998	73.37	0.1875	**73.82**	**0.1788**	+0.453	−0.00161
MoGe-2	77.14	0.1949	**79.42**	0.1873	78.28	**0.1814**	−0.006	−0.00162
UniK3D (ViT-L)	75.27	0.2652	76.17	0.2495	**76.43**	**0.2315**	+0.351	−0.01136
Depth Pro	63.32	0.2507	64.94	0.2380	**65.08**	**0.2331**	+0.312	−0.00416
UniDepth V2	70.17	0.2830	71.09	0.2789	**71.58**	**0.2608**	+0.288	−0.00535

E Additional Qualitative Comparisons

Q11 — Do the qualitative observations in Fig. 4 also hold across additional scenes?
A: Yes, the additional scenes show the same trend. Outpainting generally reduces depth errors on continuous surfaces such as walls, doors, and ceilings, with improvements visible in both camera settings (Fig. 7).

Figure 7 follows the experimental settings of Fig. 4 and provides additional comparisons between Query-only and fixed outpainting (seed 0) across five indoor and outdoor scenes.

Compared with Query-only, error reductions after outpainting generally extend across continuous surfaces such as walls, doors, and ceilings, and remain visible with known query intrinsics. These observations agree with those in Fig. 4 and are consistent with the interpretation that generated context helps mitigate scene-scale bias: additional scene context can still improve metric depth estimation even when query intrinsics are known.

F Failure Case Analysis

Q12 — Why can contextual gains in Tab. 1 fail to hold in individual scenes?
A: Visual plausibility does not guarantee reliable scale. Different generated contexts can imply different scene scales, real context need not yield more accurate predictions, and these effects vary across models.

To complement the aggregate results in Tab. 1, Fig. 8 presents three failure or counterintuitive cases, comparing the effects of real context and different generated candidates on scale estimation while keeping query pixels and model weights fixed.

In Fig. 8(a), the median predicted-to-GT depth ratios for the three generated candidates are 1.98, 0.83, and 0.64, spanning more than a threefold range; seed 0 yields median ratios above 1 for all four displayed models. This is consistent with the interpretation in Sec. B: visually plausible generated content can suggest a scale inconsistent with the actual scene, so visual quality alone does not establish its usefulness for depth estimation.

In Fig. 8(b), DA3 has an AbsRel of 0.43 with real context, compared with 0.10 with generated context (seed 0), and its scale estimate is also farther from ground truth. This shows that context authenticity does not guarantee more accurate scale estimation: real surroundings need not provide more useful scale cues than generated content.

In Fig. 8(c), Depth Pro's AbsRel rises from 0.11 with Query-only to 0.52 with real context, while the median scale ratio falls from 0.87 to 0.47. This shows that even real context can shift an initially relatively accurate prediction toward an incorrect scene scale, so context-induced degradation cannot be attributed solely to the lack of realism in generated content.

These cases illustrate limitations of contextual guidance rather than their prevalence; aggregate performance is reported in Tab. 1.

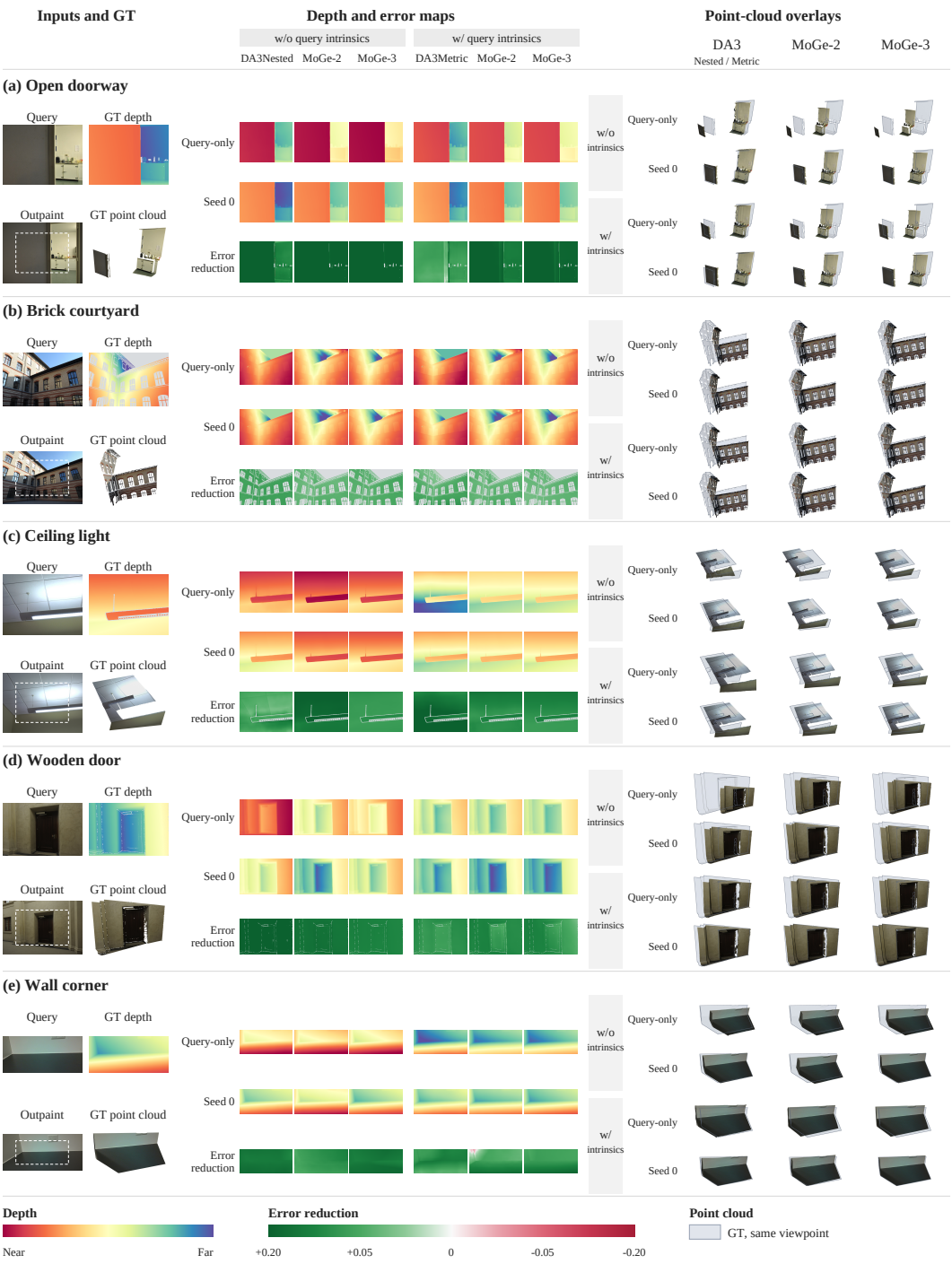

Figure 7: **Additional qualitative comparisons across five scenes.** Each row presents one scene. Left: query RGB, GT depth, seed-0 outpainting, and GT point cloud; dashed boxes mark the query ROI. Middle: Query-only depth, seed-0 depth, and error reduction, with columns grouped by query-intrinsic availability. Right: point clouds in three model columns; upper/lower groups compare Query-only and seed 0 without/with query intrinsics. DA3 uses Nested/Metric, respectively. Depth colors run from red (near) to purple (far). Green/pink indicate lower/higher per-pixel AbsRel. Light gray point clouds show GT overlaid from the same viewpoint.

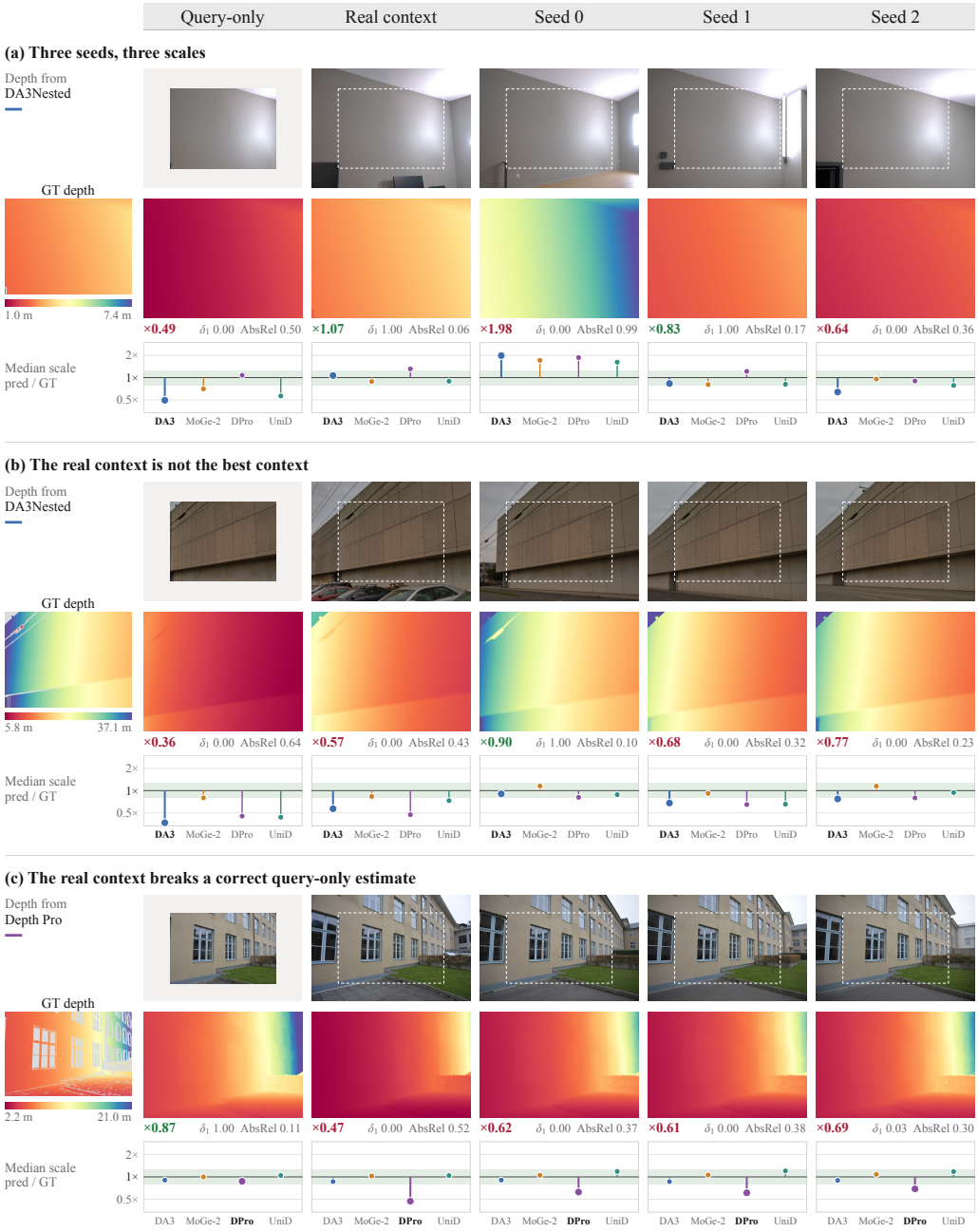

Figure 8: **Context-induced scale failures and counterintuitive cases.** Columns compare Query-only, real context, and seeds 0/1/2 from left to right. RGB inputs appear above the corresponding ROI depth predictions, with GT depth on the left. Dashed boxes mark the fixed query ROI; all metrics use its valid pixels. Predictions and GT share a depth color scale in meters within each case. The $\times$ value is the median predicted-to-GT depth ratio, with $1\times$ indicating scale agreement; δ_1 is reported as a fraction in $[0, 1]$. Bottom strips show median scale ratios for four models: colored markers correspond to the labeled models, with stems extending from $1\times$ to each ratio. The pale green band marks $[0.8, 1.25]$; triangles indicate values beyond the axis range. DA3, DPro, and UniD denote Depth Anything 3 (Nested), Depth Pro, and UniDepth V2; bold labels identify the displayed depth model. All results are without query intrinsics.