REFINE ANY DEPTH

Anonymous authors
Paper under double-blind review

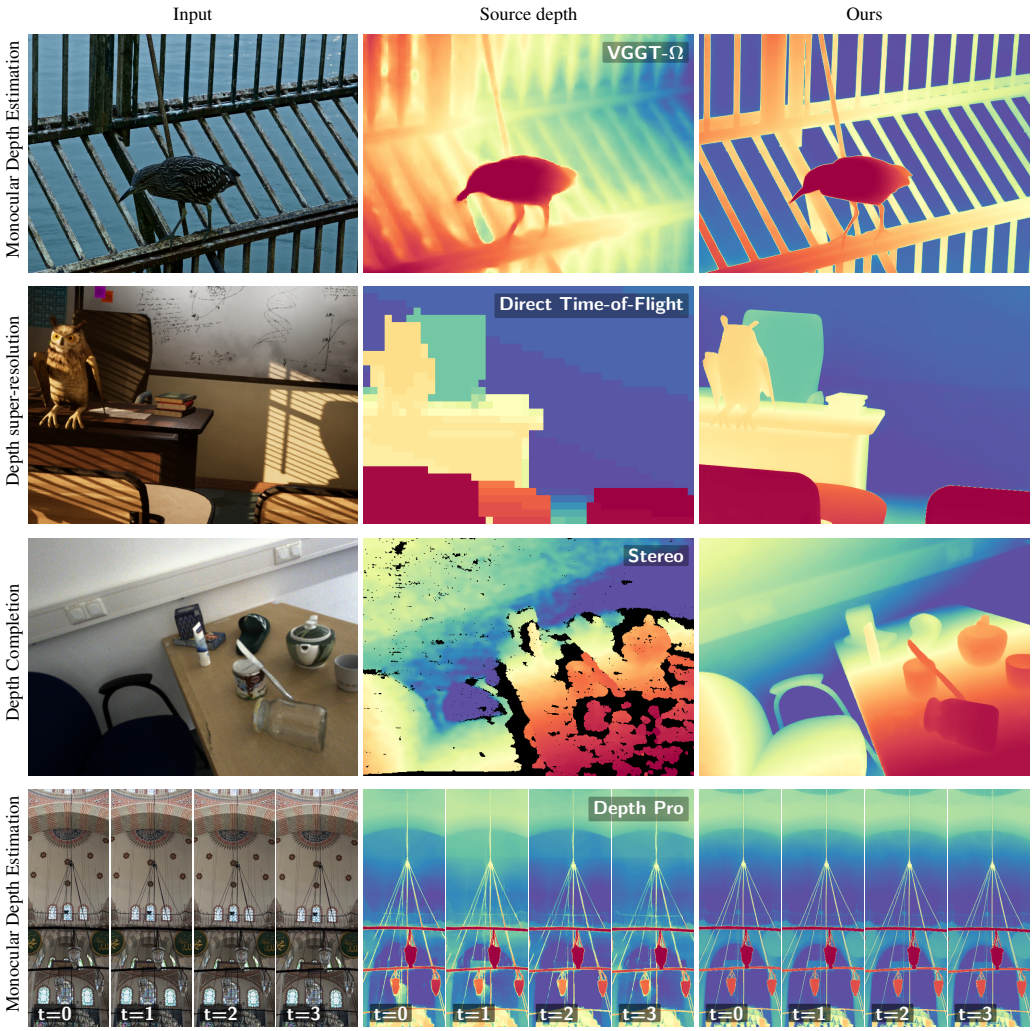

Figure 1: We present **Refine Any Depth (RAD)**, a universal depth refinement method that improves depth estimates from varying depth sources. From top to bottom, RAD sharpens boundaries in VGGT-Ω predictions (Wang et al., 2026a), recovers fine details from low-resolution direct time-of-flight measurements, completes missing regions in stereo depth, and improves scale consistency across per-frame Depth Pro predictions (Bochkovskii et al., 2025) on videos, all with a single model.

ABSTRACT

Depth maps are essential to applications such as novel view synthesis and robotics, but the depth maps produced by sensors and neural networks are frequently flawed: low resolution, noisy, incomplete, spatially inaccurate, or temporally unstable. Many existing fixes are task-specific, tailored to a single problem such as completion or super-resolution, and do not generalize beyond that setting. We introduce a unified model to refine a broad class of depth inputs from real sensor data to the outputs of state-of-the-art depth models. We formulate depth refinement as residual regression. Given an input video and depth map, we fine-tune a

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.

video diffusion model to perform a single-step residual update to repair the input depth. Training on a curriculum of depth corruptions enables generalization and strong performance across depth modalities and error types. We present a simple recipe that converges quickly with a modest amount of training data. Our model achieves state-of-the-art performance in depth super-resolution and depth completion, and improves boundary sharpness and temporal stability of state-of-the-art depth models, including lifting single-frame estimates to coherent video depth.

1 INTRODUCTION

Depth maps come from many sources, each prone to its own failure modes, and choosing among them involves trade-offs. Real-world sensors such as stereo and time-of-flight cameras provide metric measurements that are valuable for robotics, but them also produce noisy and incomplete readings which complicate tasks like manipulation or data annotation. Learned depth models have their own limitations. Some of the strongest recent models operate on single images (Gabeur et al., 2026), but their lack of temporal context can lead to frame-to-frame jitter when applied to video (Bochkovskii et al., 2025). Multiview models exploit information across frames, but may overlook fine geometric details (Wang et al., 2026a;b).

Depth refinement offers a way around these trade-offs. A depth refiner takes an existing depth map and corrects its errors, ideally addressing issues regardless of source or failure mode. A good refinement system should inpaint missing regions, upsample coarse geometry, and stabilize depth across time. These corrections have historically been studied as separate problems with largely independent literatures (Hu et al., 2022; Sun et al., 2023), but models today are capable enough to support the more ambitious goal of tackling all at once.

In this paper, we introduce a model that refines any depth (RAD). RAD can denoise and complete raw sensor data, and sharpen and stabilize depth model outputs. We formulate this as a general problem of RGB-guided residual regression. Given an input video, depth sequence, and optional confidence mask, the model predicts a correction to its input depth. We train RAD by fine-tuning a pretrained video diffusion model (Wan Team et al., 2025) with a one-step flow-matching objective. Our flow begins at the input depth rather than at noise and ends at the clean depth, so one integration step is equivalent to a residual update. Inference is a single deterministic forward pass.

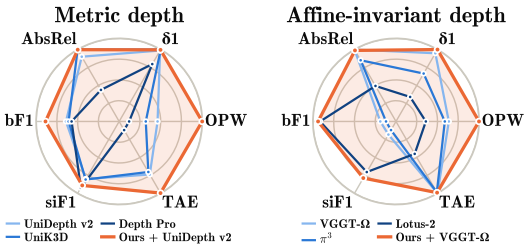

Figure 2: **RAD improves state-of-the-art metric and affine-invariant depth estimation.** We feed UniDepth v2 into our refiner for metric depth (left) and VGGT-Ω for affine-invariant depth (right). In zero-shot evaluation, the resulting predictions (orange) outperform the compared state-of-the-art methods (blue) in mean rank in both settings (Table 1). The radar charts summarize depth accuracy, boundary sharpness, and temporal consistency over selected evaluation datasets. Larger radii indicate better performance after per-metric normalization. See Appendix B.1 for details.

Because our model conditions on an existing depth estimate, it does not need to concern itself with recovering metric scale or globally accurate geometry. The input depth provides that grounding. This lets the video prior contribute sharp, well-localized boundaries, robustness across diverse real-world scenes, and temporal coherence. These properties emerge from the diversity and scale of video data seen during diffusion pretraining, and a modest amount of fine-tuning on a few depth datasets is enough to align these properties with the residual correction task.

A growing body of work conditions depth models on an existing depth estimate. Prompt Depth Anything (Lin et al., 2025) and Depth Anything with Any Prior (Wang et al., 2026c) incorporate sensor measurements or other metric priors into monocular estimators, while BetterDepth (Zhang et al., 2024), SharpDepth (Pham et al., 2025), and Marigold-DC (Viola et al., 2025) use diffusion priors to refine predictions or complete sparse depth. These efforts reflect a converging view that strong pretrained priors combined with an input depth signal are an effective recipe. However, each

operates on single images and targets a subset of tasks. We extend this direction to video and show that a single model can handle a broad range of depth refinement tasks and outperform earlier task-specific methods.

Specifically, we model imperfect depth as drawn from a mixture of corruption sources (missing measurements, noise and outliers, reduced resolution, temporal flicker), and train a network to undo the resulting corruption. The result is a simple, fast-converging recipe that generalizes well across refinement tasks without any task-specific weights or architecture decisions. This model:

- improves boundary behavior and temporal stability of state-of-the-art depth models while largely preserving overall accuracy.
- achieves state-of-the-art in depth super-resolution and depth completion.
- lifts single-frame depth estimates into temporally consistent video estimates.

2 Related Work

2.1 Depth Estimation

Discriminative depth models. Many discriminative models build on pretrained visual encoders such as BEiT (Bao et al., 2021) and the DINO family (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). Single-image methods include Depth Anything V2 (Yang et al., 2024), Depth Pro (Bochkovskii et al., 2025), UniDepth v2 (Piccinelli et al., 2025b), and MoGe-2 (Wang et al., 2025b), while UniK3D (Piccinelli et al., 2025a) extends monocular metric 3D estimation across camera models. More recent methods improve local geometry through sparse 3D refinement in MoGe-3 (Kong et al., 2026) and arbitrary-resolution prediction with neural implicit fields in InfiniDepth (Yu et al., 2026). However, independent framewise prediction does not enforce temporal consistency and can introduce flicker in videos.

Beyond single-image prediction, DUSt3R (Wang et al., 2024b) estimates pairwise point maps, while VGGT (Wang et al., 2025a), VGGT-Ω (Wang et al., 2026a), and π^3 (Wang et al., 2026b) jointly recover geometry across multiple views. Depth Anything 3 (Lin et al., 2026) unifies single- and multi-view prediction, while Video Depth Anything (Chen et al., 2025) and FlashDepth (Chou et al., 2025) introduce temporal modeling for long videos and high-resolution streaming. Despite these advances, fine structures and sharp depth boundaries remain challenging. Our approach uses a video diffusion prior to refine both spatial detail and temporal consistency in existing predictions.

Diffusion-based depth models. Early methods generate depth by denoising random noise conditioned on input images (Duan et al., 2024; Saxena et al., 2023). Subsequent approaches adapt pretrained image generators to recover finer geometric detail (Ke et al., 2024; Fu et al., 2024; Gui et al., 2025), with video extensions targeting temporally consistent depth prediction (Hu et al., 2025; Shao et al., 2025; Yang et al., 2025). A complementary line repurposes generative backbones for single-step or deterministic prediction (Bai & Huang, 2025; He et al., 2025b; Xu et al., 2025). This includes Lotus-2 (He et al., 2025a), which combines a deterministic image predictor with optional detail refinement, and DVD (Zhang et al., 2026), which performs single-pass video depth regression. Although diffusion priors can produce sharp depth maps, visual detail does not guarantee accurate scene geometry or metric scale. Our approach conditions refinement on existing depth maps to retain their geometric and scale information.

2.2 Depth Refinement

Real-world depth sensors can produce incomplete measurements, as in stereo and indirect time-of-flight (iToF) systems, or low-resolution depth, as in compact direct-ToF sensors (Jung et al., 2023; Sun et al., 2023). Depth completion methods such as OMNI-DC (Zuo et al., 2025), Any2Full (Zhou et al., 2026), and Lingbot Depth (Tan et al., 2026) recover dense depth from partial measurements, while Marigold-DC (Viola et al., 2025) and DepthLab (Liu et al., 2024) use diffusion priors for completion. CDMs (Liu et al., 2026) additionally targets camera-specific measurement errors. For low-resolution depth, DORNet (Wang et al., 2025c) models unknown degradations, while DuCos (Yan et al., 2025) and PromptDA (Lin et al., 2025) incorporate pretrained depth priors to recover detailed

geometry. DVSR (Sun et al., 2023) and STDNet (Wang et al., 2026d) exploit temporal information for video depth super-resolution. G2-MonoDepth (Wang et al., 2024a) and PriorDA (Wang et al., 2026c) further unify multiple restoration tasks and input patterns. For imperfect model predictions, BetterDepth (Zhang et al., 2024) and SharpDepth (Pham et al., 2025) use diffusion priors to improve spatial detail, while NVDS (Wang et al., 2023) stabilizes framewise depth estimates. Our universal framework spans sensor restoration and model-prediction refinement, using a pretrained video diffusion prior and residual regression to correct spatial and temporal depth errors.

3 METHOD

Our framework refines sensor measurements and model predictions using a single network initialized from a pretrained video diffusion model. The input depth from sensor or model supplies an initial geometry estimate, while the RGB video and optional addition confidence or valid mask information guide its correction. We first provide a formal definition of the task, then describe the latent refinement network and its residual-regression objective.

3.1 DEPTH REFINEMENT AS RESIDUAL REGRESSION

Let $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times 3}$ be an RGB video and $\mathbf{D}_{\text{raw}} \in \mathbb{R}_+^{T\times H\times W}$ an imperfect depth sequence aligned with its frames. Here, T, H, and W denote the number of frames, height, and width. We denote the corresponding true depth sequence by $\mathbf{D}_{\text{true}} \in \mathbb{R}_+^{T\times H\times W}$. Additionally, if confidence or validity information is available, we represent it by a map $\mathbf{C} \in [0,1]^{T\times H\times W}$. If $\mathbf{C}$ unavailable, we write $\mathbf{C} = \perp$. The goal of depth refinement is to recover $\mathbf{D}_{\text{true}}$ from condition $\mathbf{S} = (\mathbf{V}, \mathbf{D}_{\text{raw}}, \mathbf{C})$.

Under squared ℓ_2 loss, the optimal depth estimate is the conditional mean. Because the input depth is part of $\mathbf{S}$, this estimate can be expressed as the input plus an expected correction:

$$\mathbb{E}[\mathbf{D}_{\text{true}} \mid \mathbf{S}] = \mathbf{D}_{\text{raw}} + \mathbb{E}[\mathbf{D}_{\text{true}} - \mathbf{D}_{\text{raw}} \mid \mathbf{S}]. \tag{1}$$

The second term is the expected residual needed to correct the observed depth. A neural refiner parameterized by θ can therefore be written in residual form:

$$\hat{\mathbf{D}}_\theta = \mathbf{D}_{\text{raw}} + \mathbf{r}_\theta(\mathbf{S}), \tag{2}$$

where $\mathbf{r}_\theta(\mathbf{S})$ is the network's predicted depth correction. This formulates sensor measurements, model predictions, and their refinement.

3.2 MODELING DEPTH CORRUPTION

We describe imperfect depth and its confidence map using a mixture of corruption sources:

$$(\mathbf{D}_{\text{raw}}, \mathbf{C}) \sim q(\cdot \mid \mathbf{D}_{\text{true}}, \mathbf{V}) = \sum_{k=1}^{K} \pi_k \, q_k(\cdot \mid \mathbf{D}_{\text{true}}, \mathbf{V}), \tag{3}$$

where K is the number of corruption sources, q_k are different corruption types producing an imperfect depth input and its associated confidence map, which could be depth sensor or depth models, and $\pi_k \geq 0$ are sampling probabilities satisfying $\sum_k \pi_k = 1$. A source may exhibit one or more of the following corruption types.

- **Missing depth** q_{drop}**.** Real sensor measurements can be missing pixels because of occlusions, low texture, range limitations, or missing borders. Let $\mathbf{M} \in \{0,1\}^{T\times H\times W}$ be a sampled binary mask indicating which measurements are retained. We define

$$\mathbf{D}_{\text{raw}} = \mathbf{M} \odot \mathbf{D}_{\text{true}}, \qquad \mathbf{C} = \mathbf{M}, \tag{4}$$

  where $\odot$ denotes elementwise multiplication.
- **Measurement perturbations** q_{perturb}**.** Real sensor measurements may contain noise, systematic errors, or outliers. Let $\boldsymbol{\epsilon}$ be a sampled perturbation representing these errors, with values that preserve nonnegative depth:

$$\mathbf{D}_{\text{raw}} = \mathbf{D}_{\text{true}} + \boldsymbol{\epsilon}, \qquad \mathbf{C} = \perp, \tag{5}$$

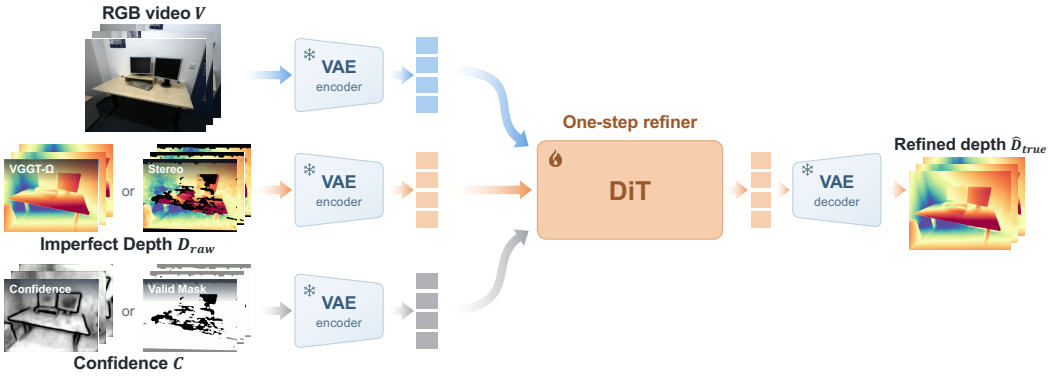

Figure 3: Our pipeline. We take either learned or sensor depth as input, and encode it to the latent space through the frozen VAE encoder. We concatenate depth tokens with clean video tokens and go through DiTs to predict clean depth tokens. The final output is decoded into the pixel space.

- **Low-resolution depth** q_{lowres}**.** Compact direct-ToF sensors provide coarse depth measurements. Let $\mathcal{O}$ be a sampled degradation operator that reduces spatial information and upsamples the result to the RGB grid:
$$\mathbf{D}_{\text{raw}} = \mathcal{O}(\mathbf{D}_{\text{true}}; \mathbf{V}), \qquad \mathbf{C} = \perp, \tag{6}$$
- **Oversmooth depth along discontinuities** q_{spatial} **or temporal instability** q_{temporal}**.** It is often the behavior of a depth estimator due to model capacity, loss design, and limitation of training data. Let f be a depth predictor whose outputs may contain spatial errors, such as blurred boundaries and missing fine structures. Its depth and confidence maps are
$$\mathbf{D}_{\text{raw}}, \mathbf{C} = f(\mathbf{V}), \qquad \text{or } \mathbf{C} = \perp \text{ if not produced by } f. \tag{7}$$

3.3 Depth Residual Regression as One-Step Flow Matching

The corruption types above call for complementary corrections: *inpainting*, *upsampling*, *spatial detail recovery*, and *temporal stabilization*. To this end, we propose to exploit pretrained video diffusion models who naturally owns strong priors on such domains due to massive pretraining.

We adopt Wan (Wan Team et al., 2025), an open-source video diffusion model, as the backbone of our residual regression network $\mathbf{r}_{\boldsymbol{\theta}}(\mathbf{S})$ to exploit its strong priors. As Wan operates in the latent space, we follow the common practice and convert RGB video $\mathbf{V}$, depth video $\mathbf{D}_{\text{raw}}$ and confidence map $\mathbf{C}$ if available into latent space. If $\mathbf{C}$ not available, we pass constant map to this branch. We independently encode them using the VAE encoder, obtaining their latent representations $\mathbf{x}_{\mathbf{V}}$, $\mathbf{x}_{\mathbf{D}_{\text{raw}}}$, and $\mathbf{x}_{\mathbf{C}}$, and concatenate them together to form $\mathbf{x}_{\mathbf{S}}$ pass in full condition $\mathbf{x}_{\mathbf{S}}$ into the DiT.

The pretrained VAE provides a compact, decodable representation of the depth signal, allowing us to use latent regression as a surrogate for depth-space regression and express the predicted endpoint in residual form:
$$\hat{\mathbf{x}}_\theta = \mathbf{x}_{\mathbf{D}_{\text{raw}}} + \mathbf{r}_\theta^{\text{lat}}(\mathbf{x}_{\mathbf{S}}), \tag{8}$$
where $\mathbf{r}_\theta^{\text{lat}}(\mathbf{x}_{\mathbf{S}})$ is the predicted depth correction in latent space. The DiT is fine-tuned to predict the refined latent representation $\hat{\mathbf{x}}_\theta$, which is subsequently decoded into a refined depth map $\hat{\mathbf{D}}_\theta$ in normalized space. When metric depth is required, the normalized depth map is affine-aligned with the input imperfect depth to recover metric scale and shift. The pipeline is shown in Fig. 3.

The latent update shown in Eq. (8) admits a rectified-flow interpretation. In the flow matching perspective, consider the linear path from the imperfect noisy depth latent to the clean target latent:
$$\mathbf{x}_t = (1-t)\mathbf{x}_{\mathbf{D}_{\text{raw}}} + t\mathbf{x}_{\mathbf{D}_{\text{true}}}, \qquad t \in [0,1]. \tag{9}$$
Its target velocity is the constant latent displacement $\dot{\mathbf{x}}_t = \mathbf{x}_{\mathbf{D}_{\text{true}}} - \mathbf{x}_{\mathbf{D}_{\text{raw}}}$, which is the residual predictor $\mathbf{r}_\theta^{\text{lat}}(\mathbf{x}_{\mathbf{S}})$ need to estimate. The update in Eq. (8) thus takes the form of a single forward Euler step with unit step size. Following prior work on single-step adaptation of diffusion models (He et al., 2025b; Zhang et al., 2026), we train our model to directly regress this latent displacement.

3.4 TRAINING

Constructing corrupted training inputs. During training, we start with pretrained Wan weights and use synthetic datasets with rendered ground-truth depth to fine-tune the DiT. We aim to cover $q(\cdot \mid \mathbf{D}_{\text{true}}, \mathbf{V})$ as we instantiate the corruption types in Section 3.2 through four training sources:

- Inspired by the study of incomplete and inaccurate sensor measurements (Tan et al., 2026; Jung et al., 2023), we sample masks that mimic occlusions and missing borders, then corrupt retained measurements with noise, covering both q_{drop} and q_{perturb}. We assign a constant confidence value to all retained measurements including corrupted ones, and zero to missing pixels.
- Following histogram-based dToF simulation in prior work (Sun et al., 2023; Wang et al., 2026d), we synthesize q_{lowres} by aggregating ground-truth depths into local histograms weighted by return strength. Each block is assigned the depth of its dominant bin and replicated onto the RGB grid. We sample downsampling factors from $\{4, 8, 16\}$ and use a constant confidence map.
- We apply frozen VGGT-Ω (Wang et al., 2026a) to $\mathbf{V}$ and use its depth predictions and confidence maps to target q_{spatial}, including blurred boundaries and missing fine structures.
- We apply frozen Depth Pro (Bochkovskii et al., 2025) independently to each frame, exposing the refiner to q_{temporal} through frame-to-frame inconsistencies. We use a constant confidence map because the predictor does not provide one. These predictions may also contain spatial errors.

Loss Functions. We supervise the refined depth latent using mean-squared error (MSE), which is equivalent to the one-step flow matching objective:

$$\mathcal{L}_{\text{fm}} = \mathbb{E}\left\|\mathbf{r}_\theta^{\text{lat}}(\mathbf{x}_{\mathbf{S}}) - (\mathbf{x}_{\mathbf{D}_{\text{true}}} - \mathbf{x}_{\mathbf{D}_{\text{raw}}})\right\|_2^2 = \mathbb{E}\left\|\hat{\mathbf{x}}_\theta - \mathbf{x}_{\mathbf{D}_{\text{true}}}\right\|_2^2. \tag{10}$$

Following DVD (Zhang et al., 2026), we apply a 3D gradient matching loss to discourage over-smoothing of spatial boundaries and temporal variations. Here, the three dimensions correspond to time, height, and width in the latent video. For each direction $d \in \{t, h, w\}$, let ∇_d denote the first-order finite-difference operator between adjacent positions along that dimension. We jointly match these differences between the predicted and target depth latents:

$$\mathcal{L}_{\text{gm}} = \sum_{d \in \{t,h,w\}} \frac{1}{N_d} \left\|\nabla_d \hat{\mathbf{x}}_\theta - \nabla_d \mathbf{x}_{\mathbf{D}_{\text{true}}}\right\|_1, \tag{11}$$

where N_d is the number of valid scalar finite differences along direction d, including all latent channels. The full training objective combines both losses: $\mathcal{L} = \mathcal{L}_{\text{fm}} + \lambda_{\text{gm}} \mathcal{L}_{\text{gm}}$.

4 EXPERIMENT

4.1 IMPLEMENTATION DETAILS

Our base model is Wan 2.1-T2V-1.3B. We freeze VAE and fine-tune its DiT backbone on TartanAir (Wang et al., 2020), Hypersim (Roberts et al., 2021), and vKITTI (Gaidon et al., 2016) for 40K steps using 8×B200 GPUs. We use the AdamW optimizer with a constant learning rate of 10^{-5} and a weight decay of 0.01. To fit depth into VAE, we simply replicate the normalized depth and confidence channels to RGB, respectively. Full training details are provided in Appendix A.

4.2 EVALUATION RESULTS

We evaluate refinement of both model predictions and sensor measurements. For model predictions, we consider both metric and affine-invariant depth, assessing accuracy, boundary quality, and temporal consistency. For sensor measurements, we evaluate depth completion and super-resolution.

4.2.1 DEPTH ESTIMATION

Refine state-of-the-art predicted depth. We refine one representative method per task, namely UniDepth v2-L (Piccinelli et al., 2025b) for metric depth and VGGT-Ω (Wang et al., 2026a) for affine-invariant depth, and compare the outcome against baselines.

Table 1: **Zero-shot depth refinement.** We use predicted camera intrinsics for metric depth refinement, and apply scale-and-shift alignment against GT for for affine-invarient depth. The rank column reports the average of the eight column ranks. Scores are scaled linearly within each column and setting, from pastel red (worst) through cream to green (best). Best scores are in **bold**. Second best underlined. Extended results can be found in Appendix B.1. All metrics except TAE and OPW are reported as percentages ($\times 100$).

Method	$\delta_1 \uparrow$ (%)				siF1↑ (%)		OPW↓	TAE↓	Rank↓
	TUM-dyn.	Bonn	Sintel	DyDToF	iBims-1	Middlebury	Sintel	7-Scenes	
Metric depth									
SharpDepth + UniDepth v2	79.1	98.0	27.1	34.2	11.3	5.44	1.34	1.77	5.50
MoGe-3	54.7	64.3	24.6	15.2	**18.8**	**14.5**	0.709	2.08	5.00
Depth Anything 3	88.1	93.8	20.3	48.3	6.42	4.06	0.581	**0.680**	4.62
Depth Pro	89.6	**98.2**	30.1	18.1	16.2	14.0	0.749	3.55	4.00
UniK3D	92.7	97.7	31.4	**56.0**	14.1	7.26	0.690	1.65	3.38
UniDepth v2	92.5	98.2	37.0	38.6	12.8	7.55	0.649	1.54	3.25
Ours + UniDepth v2	**93.3**	98.1	**38.0**	40.1	14.1	12.7	**0.492**	0.693	**2.25**
Affine-invariant depth									
NVDS + VGGT-Ω (per-frame)	81.9	79.2	75.9	83.2	1.46	0.352	0.530	1.31	9.38
PriorDA + VGGT-Ω	91.9	96.2	92.5	95.9	4.11	0.675	0.662	1.92	8.12
FlashDepth	93.1	97.7	78.2	94.1	11.1	9.26	0.683	1.70	7.88
Lotus-2	92.0	98.0	79.8	95.7	14.5	9.76	0.673	1.54	6.75
Depth Pro	92.1	98.3	83.2	94.3	16.0	10.6	0.749	3.33	6.38
DVD	88.1	98.1	78.1	95.8	13.3	3.85	0.523	0.932	6.38
Depth Anything 3	94.3	98.3	82.9	96.6	6.66	3.04	0.597	0.705	5.25
InfiniDepth	93.8	98.4	82.2	96.7	**16.1**	**15.0**	0.644	2.81	4.62
π^3	95.2	**98.6**	85.0	96.9	5.40	0.999	0.601	**0.521**	4.25
VGGT-Ω	**97.6**	97.8	96.2	97.6	4.98	0.817	0.585	0.571	4.62
Ours + VGGT-Ω	96.6	98.2	**96.4**	**97.7**	14.2	11.0	**0.470**	0.632	**2.38**

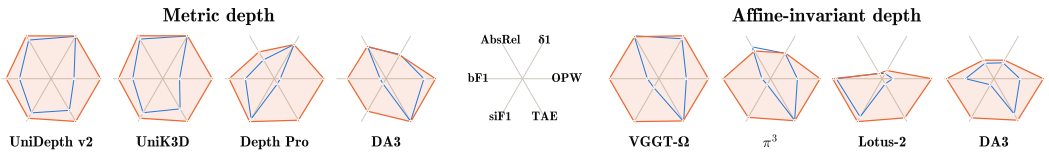

Figure 4: **RAD improves every predictor.** We feed prevalent depth model predictions into our refiner. In zero-shot evaluation, all depth models (blue) benefit from the refinement (orange). The four left panels are metric depth predictors and the four right panels are affine-invariant predictors.

The depth-estimation baselines include Depth Pro (Bochkovskii et al., 2025), UniK3D (Piccinelli et al., 2025a), MoGe-3 (Kong et al., 2026), InfiniDepth (Yu et al., 2026), Depth Anything 3 (DA3) (Lin et al., 2026), π^3 (Wang et al., 2026b), FlashDepth (Chou et al., 2025), Lotus-2 (He et al., 2025a), and DVD (Zhang et al., 2026). We additionally compare to other depth refinement methods with SharpDepth (Pham et al., 2025) applied to UniDepth v2-L, PriorDA (Wang et al., 2026c) applied to VGGT-Ω, and NVDS (Wang et al., 2023) applied to VGGT-Ω. All compared methods are evaluated zero-shot on the reported benchmarks.

We assess depth accuracy using AbsRel and δ_1, boundary quality using bF1 and siF1 (Bochkovskii et al., 2025), and temporal consistency using OPW (Wang et al., 2022) and TAE (Yang et al., 2025). Figure 2 summarizes performance across these criteria by aggregating scores over evaluation datasets, including SUN RGB-D (Song et al., 2015) for image accuracy and boundary quality. Details are provided in Appendix B.1. Table 1 provides a dataset-level comparison, reporting δ_1 on TUM-dynamics (Sturm et al., 2012), Bonn (Palazzolo et al., 2019), Sintel (Butler et al., 2012), and DyDToF (Sun et al., 2023), siF1 on iBims-1 (Koch et al., 2018) and Middlebury (Scharstein et al., 2014), OPW on Sintel, and TAE on 7-Scenes (Shotton et al., 2013). Our refined predictions achieve the best overall mean rank among the compared methods. Compared to the original predictions, refinement improves seven of eight columns for UniDepth v2 and six of eight for VGGT-Ω, with gains spanning depth accuracy, boundary quality, and temporal consistency. Our approach also outperforms other depth refinement baselines by a large margin.

Refine any predicted depth. To evaluate generalizability of our refiner across depth predictors, we further apply it to representative metric and affine-invariant models without fine-tuning. The evaluated models span single-image and video methods, including both deterministic estimators and models adapted from diffusion backbones. Figure 4 compares each predictor's original and refined outputs using the evaluation metrics and dataset aggregation described above. Every eval-

Table 2: **Zero-shot depth completion.** We report RMSE (cm), AbsRel (%) and δ_1 (%) on SCREAM. *All GT* evaluates depth of all pixels. *Interior Holes* evaluates depth in missing pixels within the sensor FoV. Scores are scaled linearly within each column, from pastel red (worst) through cream to green (best). Best scores are in **bold**. Second best underlined.

	D435						iToF					
	All GT			Interior Holes			All GT			Interior Holes		
Method	RMSE↓	AbsRel↓	δ_1 ↑	RMSE↓	AbsRel↓	δ_1 ↑	RMSE↓	AbsRel↓	δ_1 ↑	RMSE↓	AbsRel↓	δ_1 ↑
OMNI-DC	24.3	10.0	94.6	36.4	14.6	81.7	22.1	7.53	91.8	70.8	35.1	39.9
Any2Full	23.2	9.09	95.6	31.0	11.4	87.1	15.3	3.83	97.0	64.6	30.3	48.0
PriorDA	22.3	10.2	94.8	31.0	13.1	81.3	21.9	8.74	92.1	50.2	24.3	52.1
Lingbot Depth	22.4	8.90	96.0	32.1	12.1	86.1	15.8	3.77	96.3	44.3	18.1	71.0
CDMs	17.7	7.84	97.7	20.3	8.34	95.2	56.8	4.21	97.6	26.4	**10.9**	**86.5**
Ours †	**15.4**	**7.02**	**98.7**	**18.1**	**7.71**	**97.5**	**8.32**	**3.18**	**98.7**	**25.3**	11.4	85.8

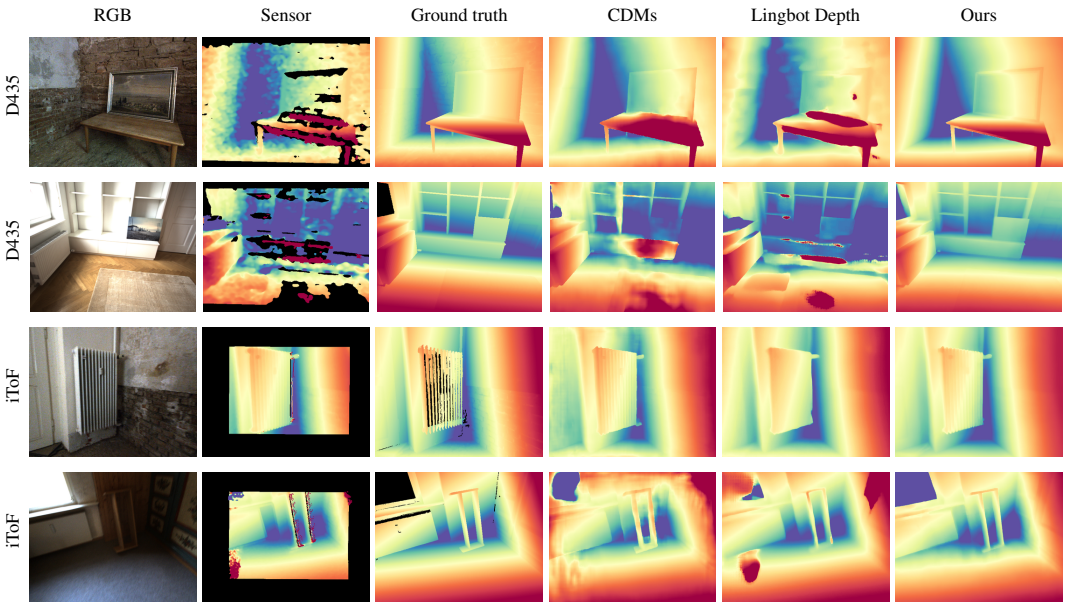

Figure 5: **Depth completion on SCRREAM.** Examples from D435 and iToF sensors. Our refinement recovers accurate geometric structure and is robust to noise.

uated predictor benefits from refinement, with particularly clear improvements in boundary quality and flow-based temporal consistency. These results demonstrate that our refiner generalizes across diverse model families, despite using predictions from only VGGT-Ω and Depth Pro during training. The improvements in spatial detail and temporal stability are consistent with our motivation for using pretrained video diffusion priors.

4.2.2 Depth Completion

Refine incomplete sensor depth. Raw RGB-D measurements can contain missing regions and inaccurate depth values. RGB-guided depth completion and restoration aim to recover dense, accurate metric depth from these degraded observations. We evaluate our method's zero-shot effectiveness on depth completion on SCRREAM (Jung et al., 2024), with synchronized real RGB video and raw depth readings from an Intel RealSense D435 active-stereo camera and a Lucid Helios indirect time-of-flight (iToF) camera. The ground-truth depth is rendered from independently scanned and registered scene meshes. We compare against representative depth completion and restoration methods, including CDMs (Liu et al., 2026), OMNI-DC (Zuo et al., 2025), Any2Full (Zhou et al., 2026), Lingbot Depth (Tan et al., 2026), and PriorDA (Wang et al., 2026c). Table 2 measures depth recovery in all pixels (*All GT*) and missing pixels in the sensor FoV (*Interior Holes*). Our method achieves best results over 10 out of 12 metrics. Figure 5 presents qualitative comparisons with CDMs and Lingbot Depth. Our predictions exhibit fewer depth artifacts, are robust to sensor noise, and better recover geometric detail. Together, the quantitative and qualitative results show that our

Table 3: **Zero-shot guided depth super-resolution.** We report RMSE (cm) and AbsRel (%) on super-resolution tasks at the scale of $4\times$, $8\times$, and $16\times$, on DyDToF and DynamicReplica. Scores are scaled linearly within each column, from pastel red (worst) through cream to green (best). Best scores are in **bold**. Second best underlined.

	DyDToF						DynamicReplica					
	×4		×8		×16		×4		×8		×16	
Method	RMSE↓	AbsRel↓	RMSE↓	AbsRel↓	RMSE↓	AbsRel↓	RMSE↓	AbsRel↓	RMSE↓	AbsRel↓	RMSE↓	AbsRel↓
Bicubic	26.20	1.00	37.16	1.43	51.00	2.25	8.83	1.35	12.22	1.66	18.10	2.33
DORNet	34.18	1.20	45.62	1.62	58.32	2.32	11.80	1.62	15.66	2.04	20.27	2.32
DuCos	38.75	1.15	45.21	1.54	59.45	2.37	13.25	1.58	15.17	1.80	21.27	2.19
PromptDA	29.95	1.46	34.13	1.54	41.89	1.67	5.86	1.35	6.05	1.34	6.85	1.38
PriorDA	24.35	0.90	29.47	1.16	33.39	1.60	6.93	0.81	10.17	1.36	11.51	1.92
STDNet	18.33	**0.68**	26.98	0.94	39.21	2.42	9.80	1.35	9.56	1.35	18.19	2.57
DVSR	22.19	0.77	23.04	**0.74**	24.99	**0.77**	6.49	1.00	6.73	0.97	7.67	1.19
Ours	**15.50**	0.70	**17.98**	0.76	**20.58**	0.78	**5.15**	**0.76**	**5.52**	**0.76**	**6.02**	**0.76**

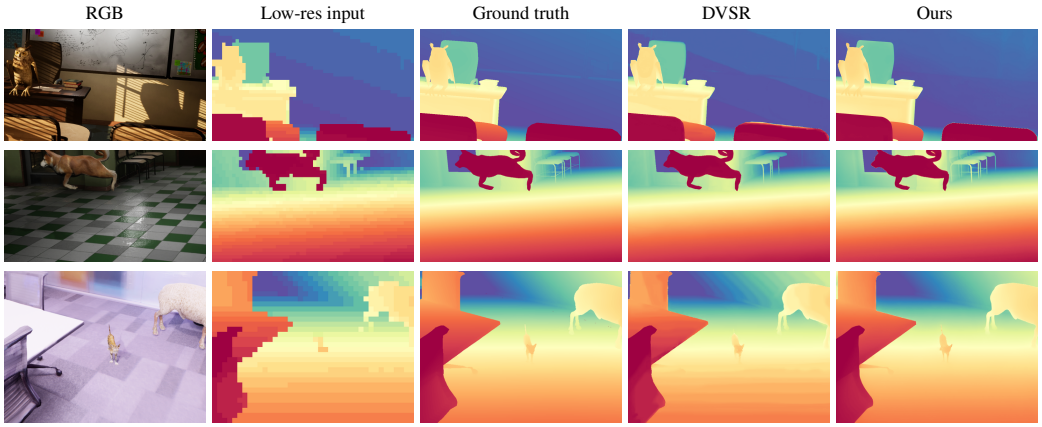

Figure 6: **Depth super-resolution at $\times 16$.** Examples from DyDToF (first two rows) and DynamicReplica (last row). Our refinement recovers finer structures and sharper object boundaries. Each row uses a shared depth color scale.

shared refiner can both recover missing depth and correct measurement errors in real sensor data, and achieve state-of-the-art results in the depth completion task.

4.2.3 Depth Super-Resolution

Refine low-resolution sensor depth. Compact dToF sensors provide metric measurements but often at much lower spatial resolution than RGB cameras (Sun et al., 2023). RGB-guided depth super-resolution aims to recover detailed depth from these coarse measurements using the complementary image information. We evaluate our refiner's zero-shot effectiveness on depth super-resolution on DyDToF (Sun et al., 2023) and DynamicReplica (Karaev et al., 2023) at $\times 4$, $\times 8$, and $\times 16$, following DVSR and STDNet (Sun et al., 2023; Wang et al., 2026d). As shown in Table 3, our model achieves the lowest RMSE and ranks first or second in AbsRel across both benchmarks and all scales. Figure 6 shows our refiner can recover sharper object boundaries and finer structures than the compared baselines. These results demonstrate that our shared refiner achieves state-of-the-art performance on depth super-resolution alongside its other refinement tasks.

5 Conclusion

We propose Refine Any Depth, a simple and effective recipe for versatile depth refinement. We model a wide range of depth corruption sources and train a network to refine imperfect depth videos through one-step flow matching. Our *single* model matches or exceeds state-of-the-art sensor depth completion and super-resolution, refines boundary sharpness and temporal stability of state-of-the-art video depth estimations, and lifts single-frame depth predictions to consistent video depth.

AI USE STATEMENT

Research and Experiments. The authors (human) come up with the original idea of the paper. We use LLM agents to conduct literature review. We use coding agents to organize and monitor experiments, and integrate open source models into our evaluation pipeline as baselines. To the best of our knowledge, we have verified the correctness of the implementation and integrations. The authors take full responsibility for the results presented in the paper.

Paper writing. The authors (human) write the paper, including equations and texts. We use coding agents to create scripts that formulate raw experiment data into LaTeX tables, and to layout figures produced by our experiments. We verify no numbers or media are hallucinated.

REFERENCES

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *CVPR*, 2022.

Yunpeng Bai and Qixing Huang. Fiffdepth: Feed-forward transformation of diffusion-based generators for detailed depth estimation. In *ICCV*, 2025.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

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 *ICLR*, 2025.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640, 2021.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, 2025.

Gene Chou, Wenqi Xian, Guandao Yang, Mohamed Abdelfattah, Bharath Hariharan, Noah Snavely, Ning Yu, and Paul Debevec. Flashdepth: Real-time streaming video depth estimation at 2k resolution. In *ICCV*, 2025.

Yiqun Duan, Xianda Guo, and Zheng Zhu. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In *ECCV*, 2024.

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 *ECCV*, pp. 241–258, 2024.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. DepthFM: Fast generative monocular depth estimation with flow matching. In *AAAI*, 2025.

Jing He, Haodong Li, Mingzhi Sheng, and Ying-Cong Chen. Lotus-2: Advancing geometric dense prediction with powerful image generative model. *arXiv preprint arXiv:2512.01030*, 2025a.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and YingCong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *ICLR*, 2025b.

Junjie Hu, Chenyu Bao, Mete Ozay, Chenyou Fan, Qing Gao, Honghai Liu, and Tin Lun Lam. Deep depth completion from extremely sparse data: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8244–8264, 2022.

Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, 2025.

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 *CVPR*, pp. 780–791, 2023.

HyunJun Jung, Weihang Li, Shun-Cheng Wu, William Bittner, Nikolas Brasch, Jifei Song, Eduardo Pérez-Pellitero, Zhensong Zhang, Arthur Moreau, Nassir Navab, et al. Scrream: Scan, register, render and map: A framework for annotating accurate and dense 3d indoor scenes with a benchmark. *Advances in Neural Information Processing Systems*, 37:44164–44176, 2024.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *CVPR*, 2023.

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.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *ECCVW*, 2018.

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.

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 *CVPR*, pp. 17070–17080, 2025.

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 *ICLR*, 2026.

Minghuan Liu, Zhengbang Zhu, Xiaoshen Han, Peng Hu, Haotong Lin, Xinyao Li, Jingxiao Chen, Jiafeng Xu, Yichu Yang, Yunfeng Lin, Xinghang Li, Yong Yu, Weinan Zhang, Tao Kong, and Bingyi Kang. Manipulation as in simulation: Enabling accurate geometry perception in robots. In *ICLR*, 2026.

Zhiheng Liu, Ka Leong Cheng, Qiuyu Wang, Shuzhe Wang, Hao Ouyang, Bin Tan, Kai Zhu, Yujun Shen, Qifeng Chen, and Ping Luo. Depthlab: From partial to complete. *arXiv:2412.18153*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. URL `https://openreview.net/forum?id=a68SUt6zFt`.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *IROS*, 2019.

Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. SharpDepth: Sharpening metric depth predictions using diffusion distillation. In *CVPR*, 2025.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3D estimation. In *CVPR*, 2025a.

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. *TPAMI*, 2025b.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021.

Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J. Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, 2014.

Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. In *CVPR*, 2025.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, 2013.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, pp. 567–576, 2015.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS*, 2012.

Zhanghao Sun, Wei Ye, Jinhui Xiong, Gyeongmin Choe, Jialiang Wang, Shuochen Su, and Rakesh Ranjan. Consistent direct time-of-flight video depth super-resolution. In *CVPR*, 2023.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception. In *ECCV*, 2026.

Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-dc: Zero-shot monocular depth completion with guided diffusion. In *ICCV*, 2025.

Wan Team, Ang Wang, Baole Ai, Bin Wen, and Chaojie Mao. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haotian Wang, Meng Yang, and Nanning Zheng. G2-MonoDepth: A general framework of generalized depth inference from monocular RGB+X data. *PAMI*, 46(5):3753–3771, 2024a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *CVPR*, 2026a.

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 *NeurIPS*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024b.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IROS*, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *ICLR*, 2026b.

Yiran Wang, Zhiyu Pan, Xingyi Li, Zhiguo Cao, Ke Xian, and Jianming Zhang. Less is more: Consistent video depth estimation with masked frames modeling. In *ACM MM*, pp. 6347–6358, 2022. doi: 10.1145/3503161.3547978.

Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *ICCV*, 2023.

Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior. In *ICLR*, 2026c.

Zhengxue Wang, Zhiqiang Yan, Jinshan Pan, Guangwei Gao, Kai Zhang, and Jian Yang. DORNet: A degradation oriented and regularized network for blind depth super-resolution. In *CVPR*, pp. 15813–15822, 2025c.

Zhengxue Wang, Yuan Wu, Xiang Li, Zhiqiang Yan, and Jian Yang. Spatiotemporal difference network for video depth super-resolution. In *AAAI*, 2026d.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? In *ICLR*, 2025.

Zhiqiang Yan, Zhengxue Wang, Haoye Dong, Jun Li, Jian Yang, and Gim Hee Lee. Ducos: Duality constrained depth super-resolution via foundation model. In *ICCV*, 2025.

Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. In *ICLR*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *NeurIPS*, volume 37, pp. 21875–21911, 2024.

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 *CVPR*, pp. 26920–26930, 2026.

Hongfei Zhang, Harold Haodong Chen, Chenfei Liao, Jing He, Zixin Zhang, Haodong Li, Yihao Liang, Kanghao Chen, Bin Ren, Xu Zheng, Shuai Yang, Kun Zhou, Yinchuan Li, Nicu Sebe, and Ying-Cong Chen. Dvd: Deterministic video depth estimation with generative priors. *arXiv preprint arXiv:2603.12250*, 2026.

Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. BetterDepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. In *NeurIPS*, volume 37, 2024.

Zhiyuan Zhou, Ruofeng Liu, Taichi Liu, Weijian Zuo, Shanshan Wang, Zhiqing Hong, and Desheng Zhang. Any to full: Prompting depth anything for depth completion in one stage. In *ECCV*, 2026.

Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. OMNI-DC: Highly robust depth completion with multiresolution depth integration. In *ICCV*, 2025.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 TRAINING DATA PREPARATION

A.1.1 TRAINING DATASETS

We train on videos from TartanAir and vKITTI, and images from Hypersim and vKITTI with synthetic perfect ground-truth depth, following Zhang et al. (2026). We sample the video datasets with relative weights of $40{:}10$ and the image datasets with relative weights of $9{:}1$. Training resolutions are 480×640 for Hypersim and TartanAir, and 352×1216 for vKITTI. For video training, we use 45-frame clips.

A.1.2 SYNTHETIC DATA CORRUPTION

We instantiate the mixture in Eq. 3 with four sources of imperfect depth, sampled with equal probability for each forward pass. Each source also supplies the confidence map $\mathbf{C} \in [0, 1]$ defined in Section 3.1, with zero reserved for missing measurements. The constant confidence values are perturbed by ± 0.03 during training.

(a) VGGT-Ω. We apply frozen VGGT-Ω (Wang et al., 2026a) to the RGB video and use its predictions to expose the refiner to spatial errors, including blurred boundaries and missing fine structures. We transform its per-pixel confidence c as $c/(c + 12)$ and clip the result to $[0.15, 0.70]$. When its confidence map is unavailable, we use a constant value of 0.55.

(b) Depth Pro. We apply frozen Depth Pro (Bochkovskii et al., 2025) independently to each frame. Its predictions provide temporal inconsistencies as well as spatial errors for the refiner to correct. Because Depth Pro does not supply confidence, we use a constant map with value 0.45.

(c) Synthetic sensor measurements. Following Lingbot Depth's strategy of combining missing depth with measurement noise (Tan et al., 2026), we corrupt ground-truth depth to simulate both missing and erroneous measurements. We extend this strategy with the following corruptions:

- Missing regions caused by occlusion shadows, low-texture dropout, or range-dependent dropout.
- A slanted band of missing depth along each image edge, with its width sampled uniformly up to 20% of the corresponding image dimension, to simulate a depth sensor with a narrower field of view than its RGB camera.
- An erroneous-depth region whose size is sampled uniformly up to 50% of the frame. Within observed regions, we shift surfaces farther away, extrapolate them, or introduce misregistration.
- A calibrated error model with a nominal mean absolute relative error of 4%, combining white noise on 20% of pixels, patch outliers on 10% at six times the gain, and heavy-tailed errors on 5%. A sampled severity parameter controls the corruption strength, ranging from clean to fully degraded depth.

We construct the confidence map by assigning a source-specific constant to observed pixels, and zero to missing pixels. Corrupted and uncorrupted observations receive the same confidence, requiring the model to use RGB evidence to distinguish accurate from erroneous measurements. During evaluation, we assign a confidence of 0.65 to valid real sensor measurements.

(d) Synthetic low-resolution depth. We instantiate the degradation using the histogram-based dToF simulation introduced by DVSR (Sun et al., 2023). Within each $s \times s$ block, we construct a histogram of ground-truth depths weighted by return strength and use the dominant bin to determine the depth measurement. We sample $s \in \{4, 8, 16\}$ and use 1024 bins over a 100 m range, yielding a depth quantization interval of approximately 9.775 cm. We replicate each measurement across its block to match the RGB grid and assign a constant confidence of 0.55.

Table 4: **Additional depth accuracy results on video datasets.** We report AbsRel and δ_1 on video datasets for the depth predictors in Figure 4 before and after refinement. Values are percentages ($\times 100$).

Method	7-Scenes		Sintel		TUM-dynamics		Bonn		DyDToF		KITTI	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Metric depth												
UniDepth v2	14.0	86.4	64.5	37.0	9.08	92.5	6.43	98.2	24.5	38.6	**6.69**	**97.6**
UniK3D	14.1	85.8	55.0	31.4	9.41	92.7	12.4	97.7	20.2	56.0	9.15	97.0
Depth Pro	11.4	89.3	70.4	30.1	12.4	89.6	6.62	98.2	39.4	18.1	14.2	82.7
Depth Anything 3	**9.46**	**94.2**	46.3	20.3	12.1	88.1	11.4	93.8	31.9	48.3	25.1	25.8
Ours + UniDepth v2	14.0	88.5	55.1	**38.0**	**8.81**	93.3	6.04	98.1	24.4	40.1	6.94	96.5
Ours + UniK3D	14.0	85.5	47.2	31.3	9.33	**94.0**	12.3	97.8	**19.2**	**59.0**	9.50	95.2
Ours + Depth Pro	11.0	90.7	53.9	35.6	11.8	89.7	**5.07**	**98.3**	41.3	9.13	13.2	83.5
Ours + Depth Anything 3	9.59	94.0	**45.7**	21.9	11.5	88.9	11.4	93.3	32.2	46.5	24.9	24.6
Affine-invariant depth												
VGGT-Ω	**5.34**	**94.8**	6.87	96.2	**3.21**	**97.6**	3.84	97.8	**2.59**	97.6	**2.63**	**98.7**
π^3	5.71	94.2	13.3	85.0	4.73	95.2	**2.85**	**98.6**	3.56	96.9	3.34	98.5
Lotus-2	6.93	93.7	18.4	79.8	8.80	92.0	4.55	98.0	5.11	95.7	6.06	95.1
Depth Anything 3	5.92	94.4	16.1	82.9	8.00	94.3	3.78	98.3	4.24	96.6	4.61	97.6
Ours + VGGT-Ω	5.79	94.3	**5.25**	**96.4**	5.25	96.6	3.72	98.2	2.88	**97.7**	3.58	98.2
Ours + π^3	5.93	94.1	12.8	85.0	6.06	94.4	3.30	98.6	3.86	97.3	4.11	98.0
Ours + Lotus-2	6.66	94.0	19.3	80.3	8.13	92.5	4.06	98.1	5.84	95.4	5.80	96.1
Ours + Depth Anything 3	6.07	94.2	14.3	83.7	8.01	93.7	3.76	98.2	4.31	96.6	4.57	97.9

A.2 Training Recipe

Forward pass pipeline. We initialize the refiner from Wan 2.1-T2V-1.3B and fix the conditioning timestep to $\tau = 0.5$ for one-step refinement. We convert the input imperfect depth video $\mathbf{D}_{\text{raw}}$ to disparity, normalize and clamp the result to $[0, 1]$. The confidence map is $\mathbf{C} \in [0, 1]^{T \times H \times W}$. When confidence is unavailable, we substitute a constant map. We supervise the model using the latent representation of the ground-truth depth video, obtained through the same encoding procedure used for the imperfect input depth video. We replicate the normalized disparity and confidence maps across three channels, then encode them and the RGB video $\mathbf{V}$ independently using the same frozen pretrained VAE. The network conditioning timestep is fixed at during training, validation, and inference.

Optimization. We optimize all DiT parameters with AdamW ($\beta = (0.9, 0.999)$, weight decay 0.01), keeping the VAE frozen. The learning rate warms up linearly from 10^{-8} to 10^{-5} over 1000 steps, then remains constant. The effective batch size is 16 clips or 128 images for single-frame training. We train in bfloat16 with gradient checkpointing and clip the gradient norm to 1.0.

Loss. We use the residual-regression loss $\mathcal{L}_{\text{fm}}$ in Eq. 10 and the latent gradient matching loss $\mathcal{L}_{\text{gm}}$ in Eq. 11, with $\lambda_{\text{gm}} = 0.5$. We compute losses only at latent positions whose corresponding ground-truth pixels are all valid. Each position corresponds to an 8×8 spatial block and, following VAE compression, one frame for the first latent frame or four frames thereafter. Gradient matching uses a finite difference only when both adjacent latent positions are valid. The temporal gradient term is zero for single-frame batches.

A.3 Inference

We refine each input clip with a single DiT forward pass and decode the refined depth latent using the frozen VAE. Clips longer than the training window are processed in 45-frame windows with 9-frame overlaps. We align consecutive windows by fitting a scale and shift in disparity space to the overlapping frames via least squares, then blend the aligned predictions in the overlap using a linear interpolation.

Table 5: **Additional depth accuracy results on image datasets.** We report AbsRel and δ_1 on image datasets for the depth predictors in Figure 4 before and after refinement. Values are percentages ($\times 100$).

Method	iBims-1		Middlebury		SUN RGB-D	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Metric depth						
UniDepth v2	8.48	**94.6**	22.2	54.9	8.55	94.7
UniK3D	**8.36**	94.4	20.1	51.2	**8.34**	**95.1**
Depth Pro	16.5	82.9	16.4	**76.7**	14.4	81.2
Depth Anything 3	10.2	86.4	22.8	49.3	8.48	93.6
Ours + UniDepth v2	8.58	94.4	22.2	54.4	8.64	94.6
Ours + UniK3D	8.48	94.2	20.1	51.3	8.45	94.9
Ours + Depth Pro	16.5	83.2	**16.3**	76.6	14.4	81.1
Ours + Depth Anything 3	10.1	86.9	22.7	47.4	8.56	93.5
Affine-invariant depth						
VGGT-Ω	3.05	97.4	2.82	98.7	3.60	97.5
π^3	2.90	97.7	2.71	99.1	**3.15**	**98.1**
Lotus-2	3.09	98.1	2.47	99.7	4.11	97.2
Depth Anything 3	3.05	97.5	2.92	99.0	3.78	97.3
Ours + VGGT-Ω	**2.82**	**98.1**	**2.31**	99.7	3.65	97.5
Ours + π^3	2.91	97.9	2.50	**99.7**	3.45	97.9
Ours + Lotus-2	3.04	98.1	2.62	99.7	4.07	97.2
Ours + Depth Anything 3	3.02	97.7	2.63	99.7	3.86	97.3

Table 6: **Additional boundary quality results.** We report bF1 and siF1 for the depth predictors in Figure 4 before and after refinement.

Method	TUM-dynamics		Bonn		DyDToF		Sintel		iBims-1		Middlebury		SUN RGB-D	
	bF1↑	siF1↑	bF1↑	siF1↑	bF1↑	siF1↑	bF1↑	siF1↑	bF1↑	siF1↑	bF1↑	siF1↑	bF1↑	siF1↑
Metric depth														
UniDepth v2	32.0	4.36	27.1	2.84	63.3	28.6	67.4	36.3	49.4	12.8	45.4	7.55	**26.7**	**3.11**
UniK3D	32.2	4.32	26.1	2.75	63.3	28.8	66.1	35.0	51.5	14.1	43.8	7.26	23.8	3.08
Depth Pro	23.7	3.30	17.6	2.18	63.4	30.1	66.0	**40.9**	54.3	**16.2**	59.8	14.0	18.9	2.20
Depth Anything 3	25.5	3.03	**27.5**	2.76	45.6	8.65	40.4	10.2	36.5	6.42	34.4	4.06	25.4	2.91
Ours + UniDepth v2	34.2	**4.74**	25.7	**3.15**	70.8	32.1	72.2	36.5	56.0	14.1	66.7	12.7	25.5	3.02
Ours + UniK3D	**34.5**	4.73	25.2	3.10	70.2	31.5	72.3	36.8	57.1	14.8	65.5	12.3	24.5	3.02
Ours + Depth Pro	31.9	4.43	22.6	2.87	**71.3**	**33.1**	**72.4**	38.2	**58.8**	16.2	**71.5**	**16.7**	23.5	2.80
Ours + Depth Anything 3	29.6	4.00	23.3	2.78	67.1	26.3	69.6	30.5	52.9	12.7	66.0	13.0	24.1	2.92
Affine-invariant depth														
VGGT-Ω	31.0	4.80	12.1	2.72	34.9	10.4	43.5	16.1	20.9	4.98	4.47	0.817	18.1	2.24
π^3	30.0	4.80	18.6	3.07	25.2	7.71	31.6	11.8	21.4	5.40	4.44	0.999	19.7	2.93
Lotus-2	34.5	4.55	25.7	**3.33**	65.2	25.6	61.9	30.2	58.6	14.5	58.0	9.76	25.5	3.19
Depth Anything 3	30.7	3.93	**28.0**	2.81	43.5	7.94	39.8	9.58	37.5	6.66	31.7	3.04	**25.6**	2.99
Ours + VGGT-Ω	35.8	5.12	21.7	2.76	**68.2**	**28.4**	**75.9**	**35.7**	54.9	14.2	60.5	11.0	24.1	2.97
Ours + π^3	**36.3**	**5.25**	24.6	3.05	59.3	22.4	65.3	28.4	52.6	12.6	56.9	9.93	25.3	**3.21**
Ours + Lotus-2	35.0	4.87	25.2	3.22	67.1	28.0	71.3	33.1	**59.0**	**14.7**	**64.6**	**12.1**	25.0	3.14
Ours + Depth Anything 3	33.3	4.63	23.9	2.86	59.6	23.0	68.0	29.5	52.6	12.7	57.8	9.94	24.2	2.94

B Evaluation Details

B.1 Depth Estimation

Radar chart aggregation. Figures 2 and 4 summarize six metrics aggregated over the following datasets:

- **AbsRel and δ_1:** Video datasets include 7-Scenes (Shotton et al., 2013), Sintel (Butler et al., 2012), TUM-dynamics (Sturm et al., 2012), Bonn (Palazzolo et al., 2019), DyDToF (Sun et al., 2023), and KITTI (Geiger et al., 2012). Image datasets include iBims-1 (Koch et al., 2018), Middlebury (Scharstein et al., 2014), and SUN RGB-D (Song et al., 2015).
- **bF1 and siF1:** TUM-dynamics, Bonn, DyDToF, Sintel, iBims-1, Middlebury, and SUN RGB-D.
- **OPW:** Sintel.
- **TAE:** 7-Scenes and KITTI.

For AbsRel and δ_1, we compute unweighted means over the video and image datasets separately, then average the two group means with equal weight. For the remaining metrics, we average scores equally across the listed datasets.

Table 7: **Additional temporal consistency results.** We report OPW on Sintel and TAE on 7-Scenes and KITTI for the depth predictors in Figure 4 before and after refinement.

Method	OPW↓ (Sintel)	TAE↓ (7-Scenes)	TAE↓ (KITTI)
Metric depth			
UniDepth v2	0.649	1.54	2.02
UniK3D	0.690	1.65	2.24
Depth Pro	0.749	3.55	3.99
Depth Anything 3	0.581	<u>0.680</u>	2.09
Ours + UniDepth v2	**0.492**	0.693	**1.62**
Ours + UniK3D	<u>0.514</u>	0.700	<u>1.63</u>
Ours + Depth Pro	0.538	0.784	1.74
Ours + Depth Anything 3	0.520	**0.654**	1.66
Affine-invariant depth			
VGGT-Ω	0.585	<u>0.571</u>	1.64
π^3	0.601	**0.521**	1.79
Lotus-2	0.673	1.54	3.86
Depth Anything 3	0.597	0.705	2.15
Ours + VGGT-Ω	<u>0.470</u>	0.632	<u>1.64</u>
Ours + π^3	0.495	0.654	**1.63**
Ours + Lotus-2	**0.465**	0.801	1.73
Ours + Depth Anything 3	0.479	0.659	1.68

Per-dataset results. We report per-dataset results for the predictors in Figure 4, before and after refinement. Table 4 reports AbsRel and δ_1 on video datasets. Table 5 reports the same metrics on image datasets. Table 6 reports bF1 and siF1. Table 7 reports OPW on Sintel and TAE on 7-Scenes and KITTI. For metric depth, we recover absolute depth by fitting a single scale and shift in disparity space over the entire clip, aligning our affine-invariant predictions with the input depth while preserving temporal consistency.

B.2 Depth Completion

Dataset. We evaluate raw depth from the D435 and Helios iToF sensors in SCRREAM, registered to the corresponding RGB frames. The evaluation set contains 16,437 D435 frames and 15,964 iToF frames. We exclude one iToF sequence because its capture is not registered to the rest of its trajectory.

Baselines. Table 2 reports results for released checkpoints of CDMs, OMNI-DC, Any2Full, Lingbot Depth, and PriorDA. For CDMs, we use the D435 checkpoint for D435 inputs and the L515 checkpoint for Helios iToF inputs. We use OMNI-DC v1.1, the ViT-L version of Any2Full, and the original pretrained ViT-L checkpoint of Lingbot Depth. All baselines perform framewise inference.

Sensor alignment. We recover metric depth by fitting a scale and shift in disparity space to align our affine-invariant predictions with observed sensor depths independently for each frame. We align per-frame because sensor gives reliable metric scale. We use ordinary least squares for D435 and trimmed least squares for Helios iToF. For the latter, we discard 20% of fitting pixels to reduce the influence of its stronger noise and outliers. No ground-truth depth is used for alignment.

Evaluation regions. We evaluate overall depth recovery and missing-depth completion in two regions:

- *All GT*: All pixels with valid ground-truth depth, including both observed and missing sensor regions. This evaluates depth recovery through both measurement correction and completion.

- *Interior Holes*: Pixels with valid ground-truth depth but missing sensor measurements within the sensor field of view (FoV). This evaluates missing-depth recovery, excluding missing borders outside the sensor FoV.

B.3 Depth Super-Resolution

Input synthesis. We follow the RGB-based peak-detection pipeline introduced by DVSR and adopted by STDNet (Sun et al., 2023; Wang et al., 2026d). Each non-overlapping $s \times s$ image region corresponds to the field of view of one low-resolution sensor pixel. Within each region, we quantize normalized ground-truth depths into $1{,}024$ bins and accumulate contributions weighted by the mean RGB intensity and inverse-square depth attenuation. The depth corresponding to the strongest histogram bin provides the low-resolution measurement. The input and target depth resolutions are $(H/s) \times (W/s)$ and $H \times W$, respectively.

Dataset and baselines. At each scale $\{4, 8, 16\}$, we evaluate 576 frames from DyDToF and 500 from DynamicReplica. Table 3 compares bicubic interpolation and DORNet, DuCos, PromptDA, PriorDA, DVSR and STDNet.

Depth alignment. As in depth completion, we recover metric depth by aligning our affine-invariant predictions to the input depth using ordinary least squares independently for each frame.

C Additional Results

C.1 Ablations

C.1.1 Data Corruption Sources

We examine how data corruption sources affect refinement of model predictions, depth completion, and depth super-resolution. We design four configurations of training sources:

- *1 Condition*: Trained only on VGGT-Ω predictions.
- *2 Conditions*: Adds Depth Pro predictions to *1 Condition*.
- *3 Conditions*: Adds low-resolution depth to *2 Conditions*.
- *4 Conditions*: Adds synthetic sensor corruption to *3 Conditions*, the full training sources.

We trained a LoRA with $\mathrm{rank} = 512$ per setup.

Table 8 evaluates refinement of model predictions using δ_1 on Sintel, siF1 on Middlebury, and TAE on Neural RGB-D (Azinović et al., 2022). We include the unrefined VGGT-Ω and Depth Pro predictions for reference. For super-resolution and completion, we report D_{105}, AbsRel, and RMSE on DyDToF at $\times 4$ and the SCRREAM iToF *All GT* region, respectively. D_{105} uses the same depth-ratio accuracy criterion as δ_1, with a threshold of 1.05 instead of 1.25. We report it as a percentage for super-resolution and a fraction for completion.

The results highlight the complementary benefits of the training sources. Training only on VGGT-Ω predictions generalizes to spatial refinement of Depth Pro outputs, consistent with both predictors exhibiting spatial errors. Temporal transfer is more limited, as the temporally consistent VGGT-Ω predictions provide fewer examples of temporal artifacts. Adding Depth Pro predictions supplies these examples, reducing TAE on Neural RGB-D from 0.727 to 0.202 when refining its outputs. Adding low-resolution depth then improves super-resolution, reducing RMSE on DyDToF from 38.05 to 15.92 cm. Finally, synthetic sensor corruption with missing and erroneous measurements improves completion, reducing RMSE on SCRREAM iToF from 0.882 to 0.133 m. Adding new corruption sources also largely preserves performance on previously covered input types, suggesting that the model has sufficient capacity to learn these complementary refinements jointly.

C.1.2 Confidence Map

We compare two models to isolate the contribution of the confidence map: one encodes only RGB and input depth, while the other additionally encodes $\mathbf{C}$ as a third latent stream. Both configurations

Table 8: **Ablation study of data corruption sources .**

	Depth Refinement (VGGT-Ω Conditioned)			Depth Refinement (Depth Pro Conditioned)			Depth Super-Res (DyDToF)			Depth Completion (SCRREAM)		
	Sintel $\delta_1 \uparrow$	Middlebury siF1↑	Neural RGB-D TAE↓	Sintel $\delta_1 \uparrow$	Middlebury siF1↑	Neural RGB-D TAE↓	$D_{105} \uparrow$	AbsRel↓(%)	RMSE↓(cm)	iToF $D_{105} \uparrow$	iToF AbsRel↓	iToF RMSE↓(m)
VGGT-Ω	**96.16**	0.82	0.167	–	–	–	–	–	–	–	–	–
Depth Pro	–	–	–	83.23	10.56	2.023	–	–	–	–	–	–
1 Condition	95.40	11.32	**0.166**	83.13	12.64	0.727	90.1	2.43	25.98	0.387	0.402	0.771
2 Conditions	95.39	10.96	0.178	83.01	**13.71**	0.202	70.4	4.65	38.05	0.385	0.407	0.783
3 Conditions	95.31	**11.38**	0.170	**83.44**	13.23	**0.188**	**97.8**	**1.07**	**15.92**	0.395	0.458	0.882
4 Conditions	95.36	10.61	0.167	83.00	12.94	0.195	97.7	**1.07**	16.73	**0.676**	**0.055**	**0.133**

Table 9: **Ablation study of confidence-map.** Both models encode RGB and input depth. The model with confidence additionally encodes the confidence map.

	Depth Refinement (VGGT-Ω Conditioned)	Depth Refinement (Depth Pro Conditioned)	Depth Super-Res (DyDToF)	Depth Completion (SCRREAM)
	Sintel AbsRel↓(%)	Sintel AbsRel↓(%)	×8 RMSE↓(cm)	D435 AbsRel↓(%)
Without confidence	20.11	32.88	34.70	2.36
With confidence	**16.46**	**30.49**	**25.36**	**2.17**

use the main model's full fine-tuning recipe, with identical corruption sources, sampling weights, and training data. We evaluate both on $10{,}000$ steps, to save computing resources but need enough to give useful signal.

Table 9 reports AbsRel (%) on Sintel with VGGT-Ω and Depth Pro inputs, RMSE (cm) on DyDToF at $\times 8$, and AbsRel (%) on SCRREAM D435. Super-resolution is evaluated in metric depth without alignment to ground truth, following Section B.3.

Adding the confidence-map stream improves all four reported scores. On Sintel, AbsRel decreases from 20.11% to 16.46% for VGGT-Ω inputs and from 32.88% to 30.49% for Depth Pro inputs. It also reduces super-resolution RMSE from 34.70 to 25.36 cm and completion AbsRel from 2.36% to 2.17%. These consistent gains across tasks support using confidence and validity cues to guide refinement of different depth sources.

C.2 Inference Runtime and Computational Cost

Table 10 reports inference latency, floating-point operations (FLOPs), multiply–accumulate operations (MACs), and parameter counts for individual components and the complete pipeline. We benchmark a fixed synthetic 45-frame clip at 480×640 resolution in bfloat16 on an NVIDIA A100-SXM4-40GB and an NVIDIA B200. Latency is averaged over 10 iterations after 3 warmup iterations on each GPU. Runtime and operation counts are reported per clip, with one multiply–accumulate counted as two FLOPs.

The *Encoder* performs three VAE encoding passes for RGB, input depth, and confidence. The *DiT* predicts the refined latent in one forward pass, and the *Decoder* converts the refined latent to depth. *Final (refiner-only)* combines these components to process an existing depth input. *Final (end-to-end)* additionally runs VGGT-Ω to generate the input depth from RGB. Both total runtimes include host overhead.

Our single-step refiner achieves an amortized throughput of 5.06 fps on A100 and 18.48 fps on B200. Including VGGT-Ω for input depth estimation, end-to-end throughput is 3.56 and 12.51 fps, respectively.

C.3 Additional Depth Estimation Qualitative Results

Figures 7–11 show five qualitative results of video depth estimation. Columns follow time from left to right. Rows show RGB input, Depth Pro (Bochkovskii et al., 2025), Depth Anything 3 (giant) (Lin et al., 2026), VGGT-Ω (Wang et al., 2026a), and our refinement of VGGT-Ω. These examples illustrate recovery of fine spatial structure across sampled video frames.

Table 10: **Inference runtime and computational cost.** Runtime is reported in milliseconds per 45-frame clip. The *Final* rows include host-device transfer overhead.

Component	Runtime (A100)	Runtime (B200)	GFLOPs	GMACs	Params
Encoder	3,584	1,165	210,277	105,138	54M
VGGT-Ω	2,716	1,161	347,530	173,765	1,144M
DiT	2,769	551	456,264	228,132	1,419M
Decoder	1,968	638	117,874	58,937	73M
Final (refiner-only)	8,887	2,435	784,415	392,208	1,546M
Final (end-to-end)	12,631	3,596	1,131,945	565,973	2,690M

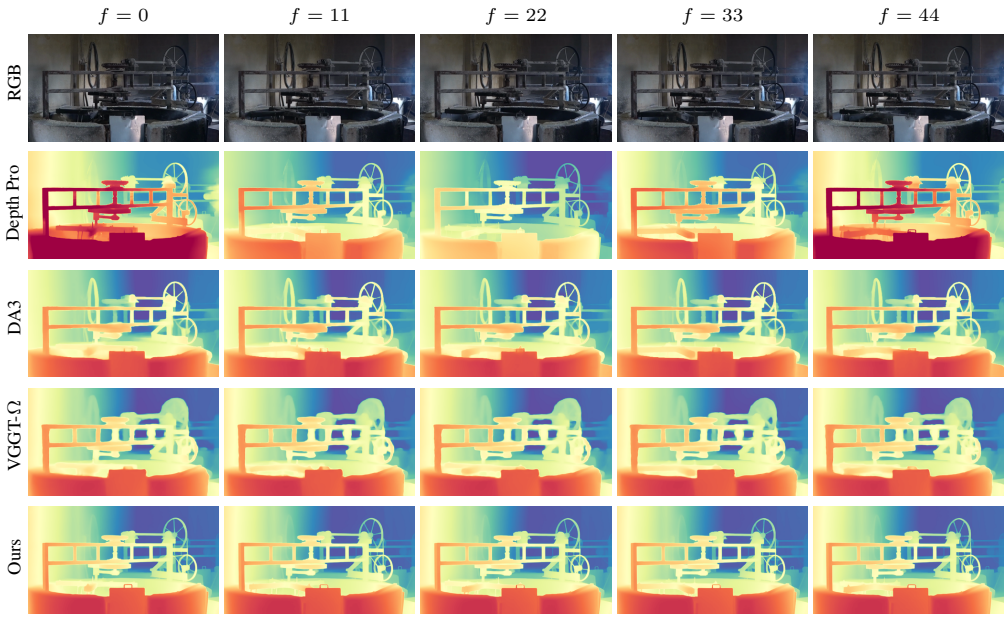

Figure 7: Additional depth estimation qualitative results.

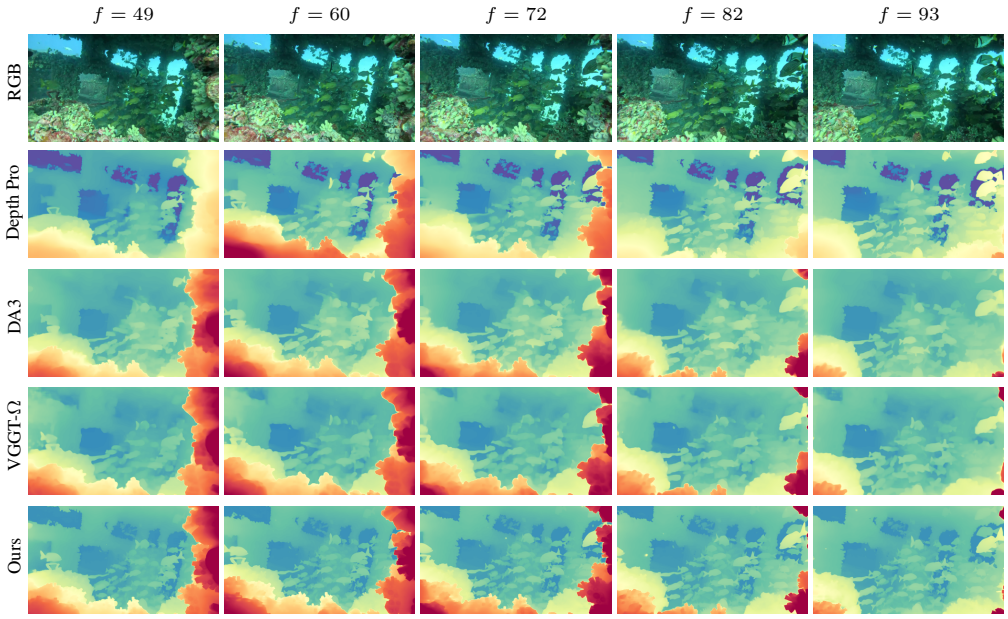

Figure 8: Additional depth estimation qualitative results.

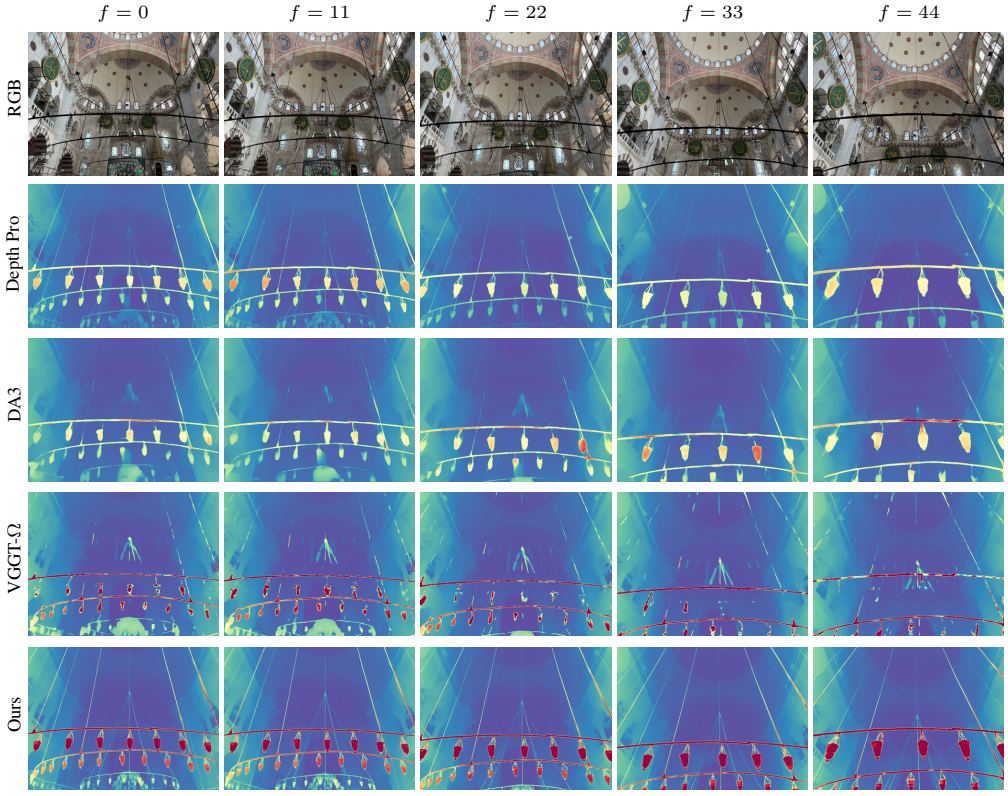

Figure 9: Additional depth estimation qualitative results.

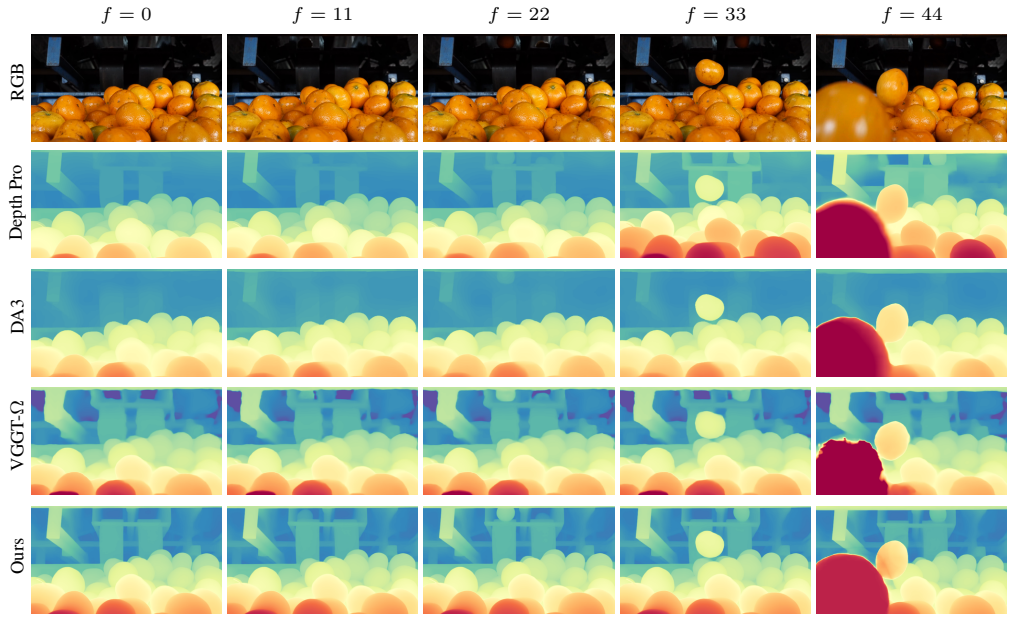

Figure 10: Additional depth estimation qualitative results.

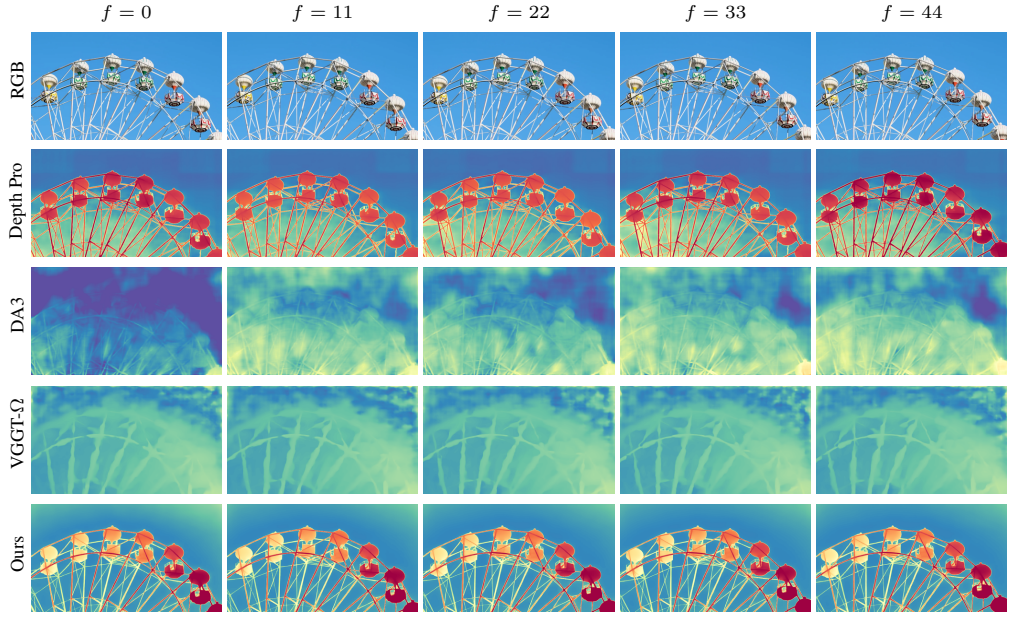

Figure 11: Additional depth estimation qualitative results.

C.4 Additional Depth Super-Resolution Qualitative Results

Figures 12–14 show selected $\times 16$ depth super-resolution results on DyDToF and DynamicReplica, comparing our refinement with DVSR and STDNet. Our refinement recovers fine geometric detail from the low-resolution inputs.

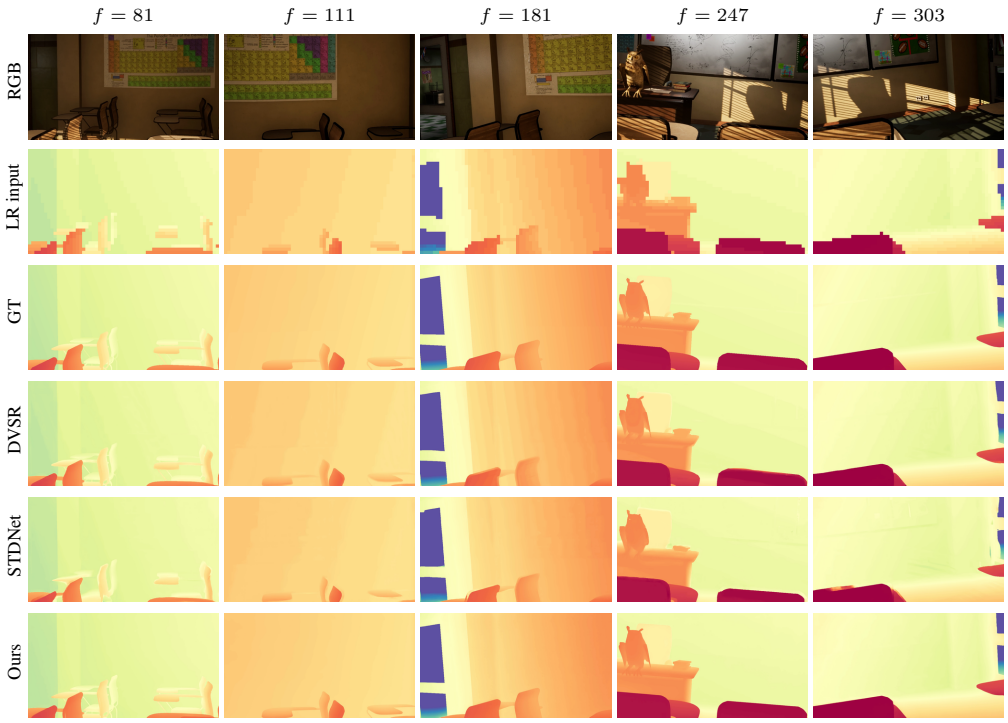

Figure 12: Additional depth super-resolution qualitative results.

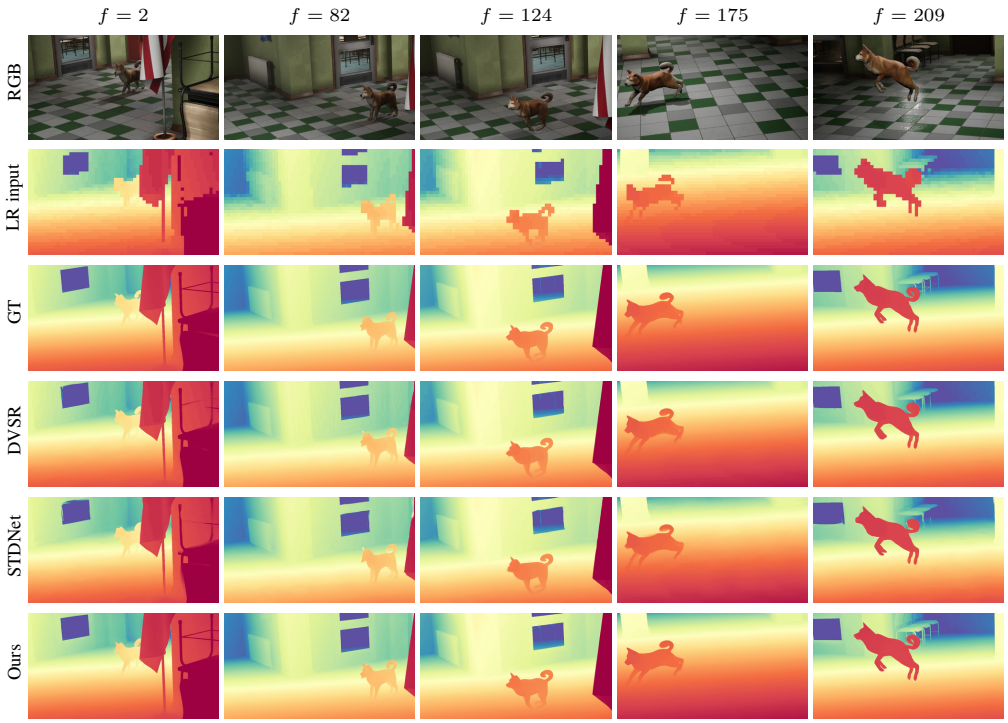

Figure 13: Additional depth super-resolution qualitative results.

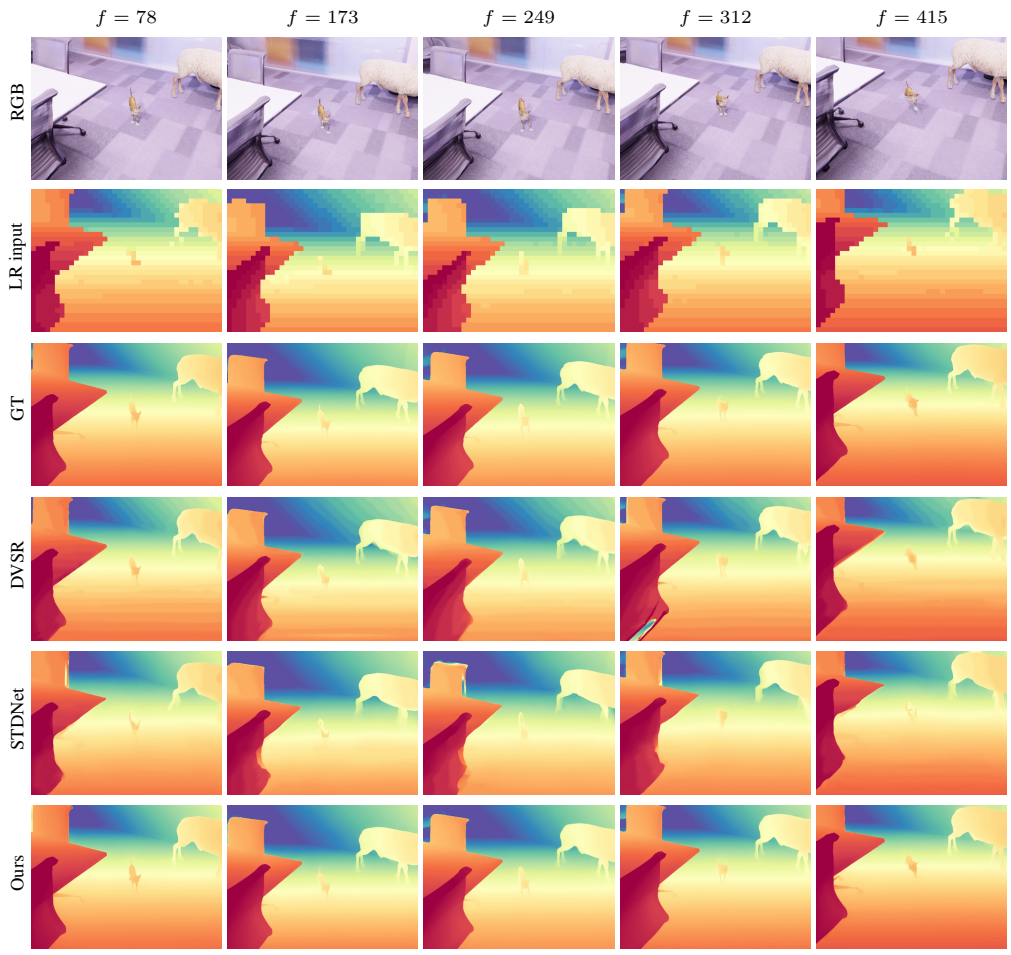

Figure 14: Additional depth super-resolution qualitative results.

C.5 Additional Depth Completion Results

Figures 15–18 show selected SCRREAM sequences recorded with D435 and iToF sensors, comparing our refinement with CDMs and Lingbot Depth. Our refinement recovers accurate geometric structure and is robust to noise.

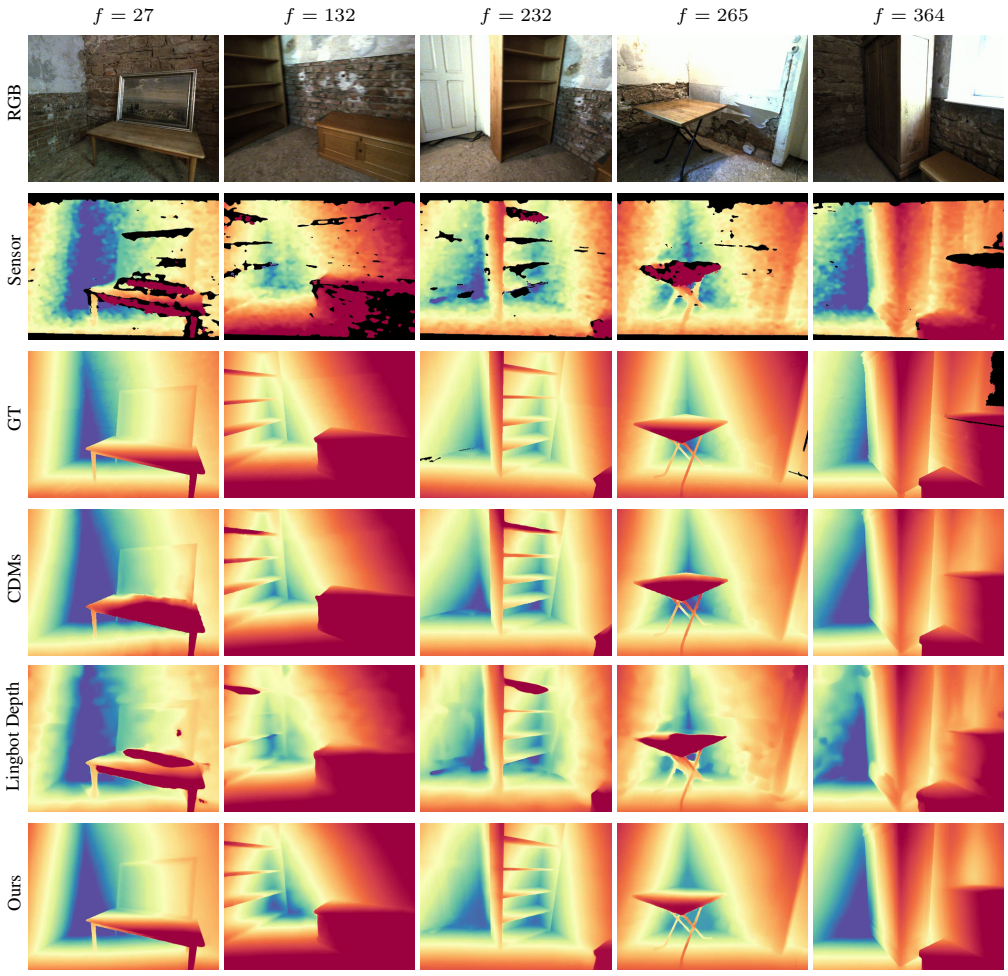

Figure 15: Additional depth completion qualitative results.

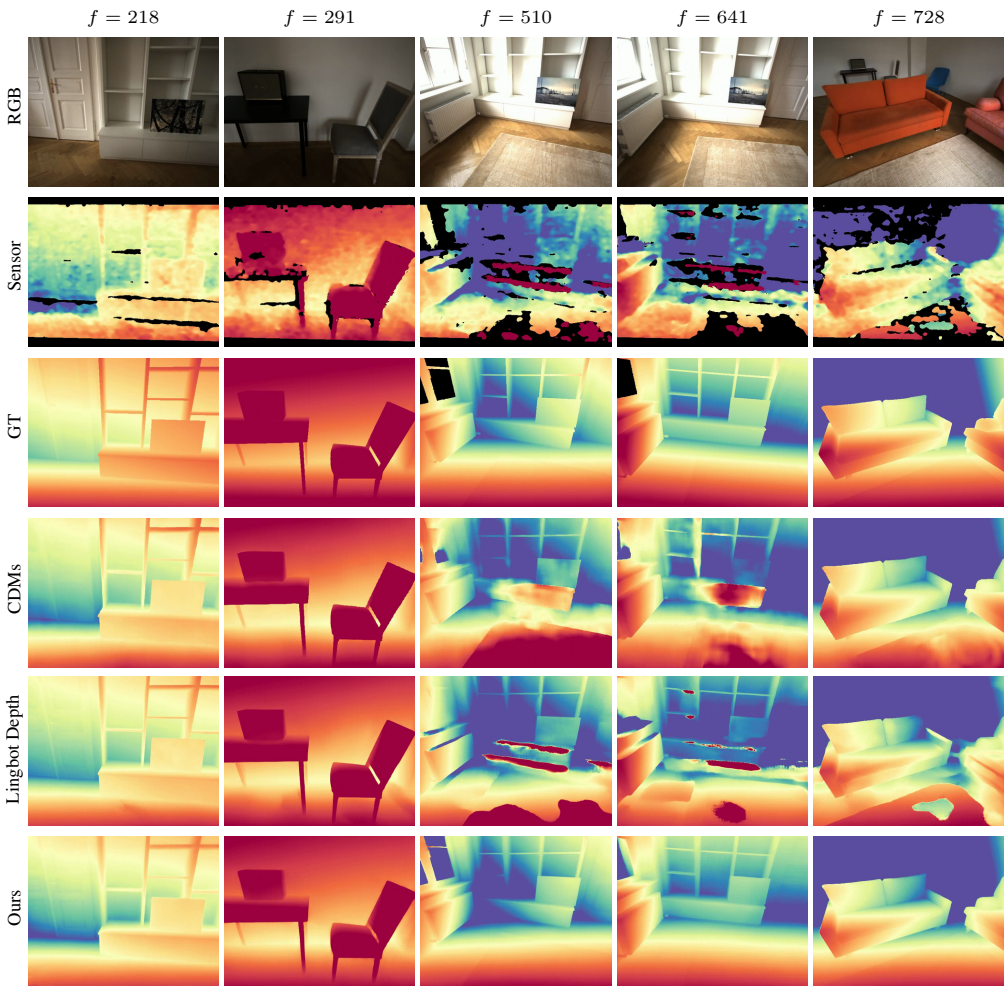

Figure 16: Additional depth completion qualitative results.

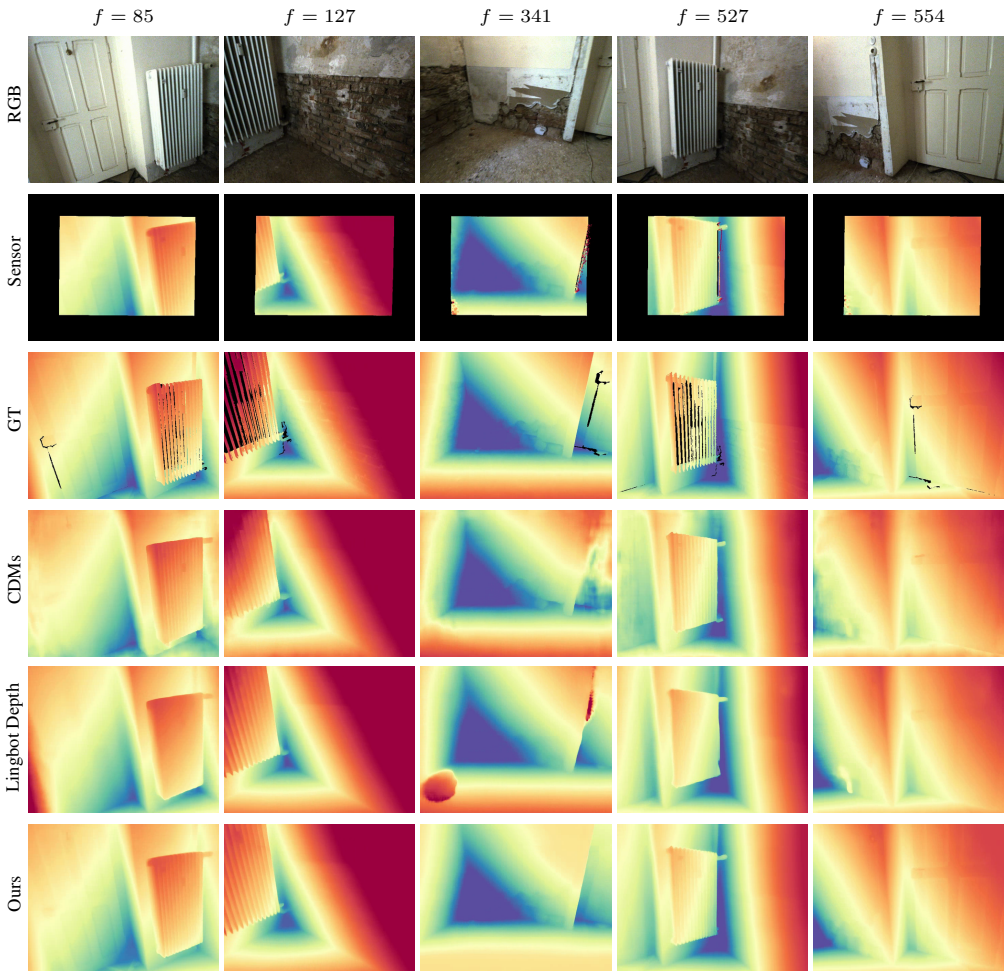

Figure 17: Additional depth completion qualitative results.

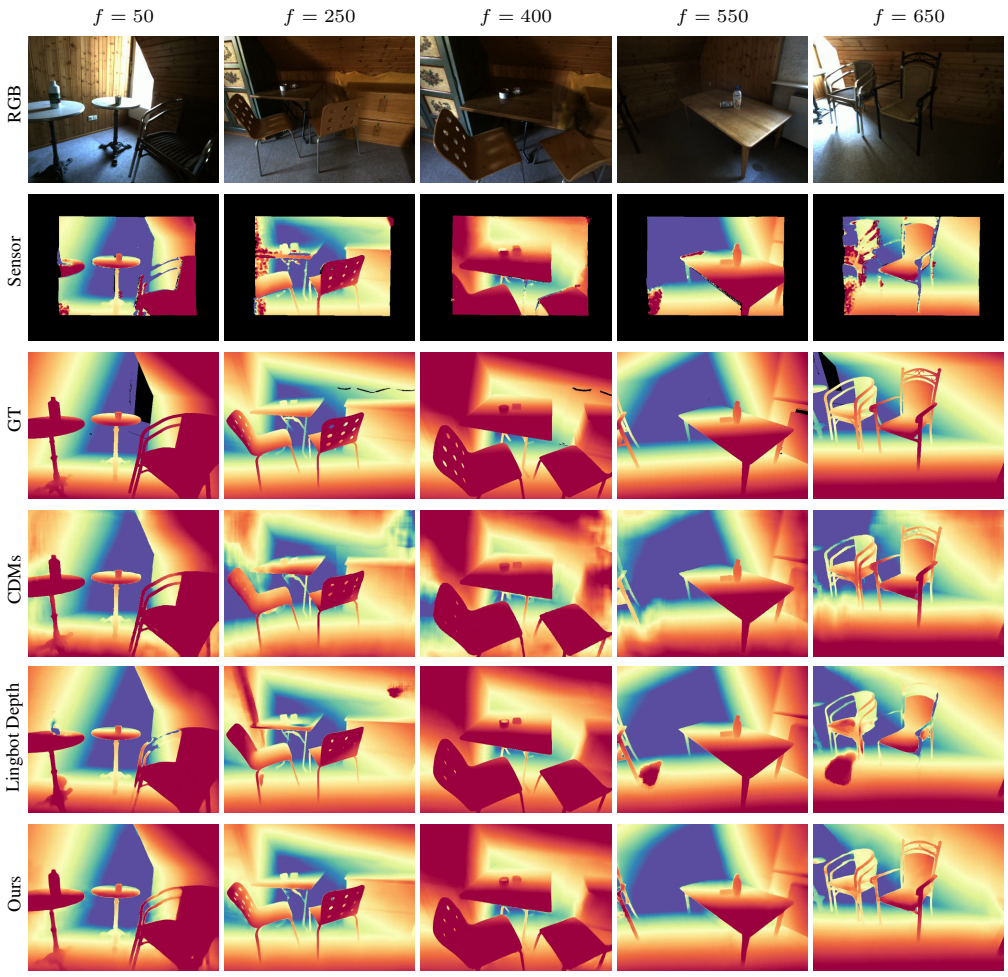

Figure 18: Additional depth completion qualitative results.

C.5.1 Depth Completion on HAMMER

Dataset. We additionally evaluate depth completion and restoration on HAMMER (Jung et al., 2023), which provides aligned RGB and raw depth from D435 active-stereo, L515 dToF, and Helios iToF cameras. We process $1{,}414$ RGB-D images per sensor independently, complementing the video evaluation on SCRREAM.

Baselines. We compare against released checkpoints of CDMs, OMNI-DC, Any2Full, Lingbot Depth, PriorDA, and Marigold-DC (Viola et al., 2025). For CDMs, we use the L515 checkpoint for all three sensors. The baselines predict metric depth directly and are evaluated without additional alignment.

Sensor alignment. We recover metric depth by fitting a scale and shift in disparity space to observed sensor measurements independently for each image. We use ordinary least squares for D435 and trimmed least squares for L515 and Helios iToF, discarding 20% of fitting pixels in the latter two cases. No ground-truth depth is used for alignment.

Results. Table 11 reports RMSE in centimeters and AbsRel over *All GT* and *Interior Holes*, as defined in Section B.2. Our method achieves the lowest errors in six of the twelve comparisons, covering both evaluation regions on L515 and *Interior Holes* on iToF, showing competitive results to state-of-the-art methods.

Figure 19 illustrates recovery of missing surfaces and correction of noisy measurements across D435, L515, and iToF inputs. In the selected examples, our refinement reduces spurious surface variations relative to Lingbot Depth, most visibly on the L515 tabletops, while preserving object boundaries.

D Limitations and Future Work

Although RAD achieves 18.48 fps on an NVIDIA B200, this throughput is amortized over 45-frame clips. Buffering and processing an entire clip introduces latency before refined predictions become available, limiting its applicability to online settings. Extending RAD to a causal, autoregressive model is a promising direction for reducing buffering latency and enabling incremental depth refinement.

Our method also inherits the reconstruction limitations of the pretrained VAE, as residual regression is performed in its latent space. Lossy encoding and decoding can introduce depth errors even when the predicted latent matches the ground-truth latent, potentially limiting geometric accuracy and the recovery of fine structures. Future work could explore depth-specific autoencoders or additional refinement in pixel space to reduce these reconstruction errors.

Table 11: **Depth completion on HAMMER.** We report RMSE (cm) and AbsRel over *All GT* and *Interior Holes*, as defined in Section B.2.

	D435				L515				iToF			
	All GT		Interior Holes		All GT		Interior Holes		All GT		Interior Holes	
Method	RMSE (cm)↓	AbsRel↓	RMSE (cm)↓	AbsRel↓	RMSE (cm)↓	AbsRel↓	RMSE (cm)↓	AbsRel↓	RMSE (cm)↓	AbsRel↓	RMSE (cm)↓	AbsRel↓
OMNI-DC	10.34	0.073	7.75	0.070	7.65	0.059	17.28	0.179	28.43	0.235	28.03	0.239
Any2Full	5.07	0.040	5.13	0.042	6.61	0.046	13.82	0.135	6.56	0.049	23.96	0.211
PriorDA	8.12	0.067	7.81	0.067	8.67	0.071	15.15	0.131	15.47	0.138	21.94	0.199
Marigold-DC	14.31	0.065	10.32	0.057	6.90	0.047	12.19	0.104	24.70	0.128	14.22	0.135
CDMs	**3.17**	**0.021**	3.75	**0.026**	3.22	0.027	6.64	0.044	5.93	0.043	10.19	0.078
Lingbot Depth	3.57	0.029	**3.60**	0.030	3.41	0.024	6.54	0.043	**3.48**	**0.032**	8.82	0.076
Ours	3.42	0.032	3.73	0.035	**2.77**	**0.019**	**5.51**	**0.034**	5.59	0.072	**7.98**	**0.069**

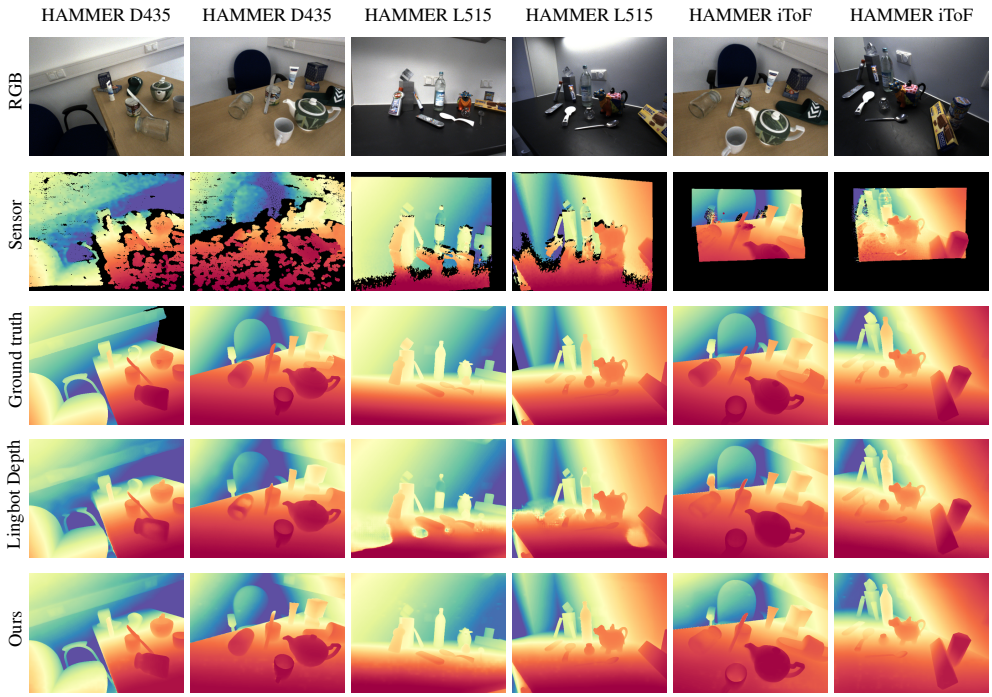

Figure 19: Qualitative depth completion on HAMMER.