Metric Depth Estimation from Arbitrarily Degraded Low-Resolution Depth Prompts

Kun Wang[1], **Yun Zhu**[2], **Pan Zhou**[3] **and Na Zhao**[1]*
[1]Singapore University of Technology and Design
[2]Nanjing University of Science and Technology [3]Singapore Management University

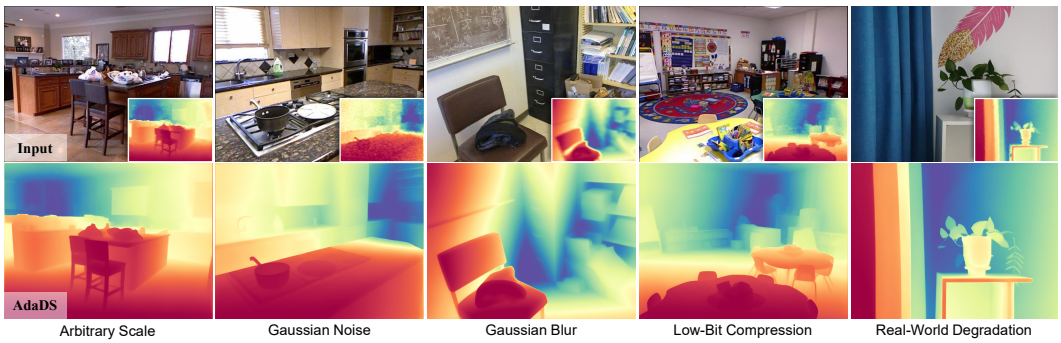


Figure 1: Given an RGB image and an arbitrarily degraded low-resolution depth prompt, **AdaDS** adaptively places the intermediate noisy sample in a high-density region of a pre-trained depth diffusion prior, enabling zero-shot estimation of fine-grained high-resolution metric depth.

Abstract

We propose AdaDS, an **Ada**ptive **D**iffusion **S**ampling framework for prompted metric depth estimation, which estimates high-resolution metric depth from an image and an arbitrarily degraded low-resolution depth prompt. This setting is commonly studied as depth super-resolution, where existing methods typically regress depth directly without an explicit prior over high-quality depth, and thus often produce artifacts under severe or unseen degradations. In contrast, AdaDS exploits the contraction property of Gaussian smoothing: as noise accumulates in the forward diffusion process, the noisy distributions of degraded prompts and their high-quality counterparts converge toward an isotropic Gaussian prior, while the prompt content is gradually erased. Accordingly, AdaDS refines the prompt and estimates its uncertainty to adaptively select a starting timestep, and injects learned noise to place the intermediate sample in a high-density region of the depth distribution learned by a pre-trained diffusion model. Denoising from this sample yields high-quality depth with fine-grained details and corrects residual errors, even under imperfect prompt refinement. Experiments on real-world and synthetic benchmarks demonstrate its superior zero-shot generalization and robustness to diverse degradations. Code is available at `https://github.com/w2kun/AdaDS`.

1 Introduction

Monocular depth estimation is fundamental to 3D perception [34, 48, 39, 53], yet it remains limited by inherent scale and geometric ambiguities. In practical settings, low-cost range sensors, such as

*Corresponding author: `na_zhao@sutd.edu.sg`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Time-of-Flight (ToF) cameras and structured-light systems, provide coarse but valuable metric depth cues. We investigate *prompted metric depth estimation*, which estimates high-resolution metric depth from an image and an arbitrarily degraded low-resolution depth prompt. This formulation subsumes conventional depth super-resolution (DSR), which typically treats the low-resolution depth as reliable and focuses on recovering high-frequency details, whereas we regard it as a geometric prompt that may contain substantial errors. By combining RGB semantics with metric cues from depth prompts, this task reduces the cost of acquiring high-quality depth data and benefits downstream applications such as 3D scene reconstruction [1, 11, 10, 40, 27] and embodied reasoning [49, 35].

Numerous methods have achieved remarkable results under controlled DSR settings, such as fixed upsampling factors, specific scene types, and limited degradation patterns [28, 36]. However, real-world depth prompts acquired from hardware-constrained sensors often exhibit complex and unpredictable degradations, making robustness a critical limitation for practical deployment. Most existing approaches directly map degraded prompts to depth [18, 38], so prompt errors propagate into their predictions and severely degrade their real-world performance.

To address this challenge, we propose AdaDS, a robust framework for prompted metric depth estimation under arbitrary depth degradation. Rather than strictly following the prompt, AdaDS projects its estimate onto a high-density region of the high-quality depth distribution learned by a pre-trained diffusion model [9, 13], allowing the generative prior to correct prompt errors. Our key insight is that the forward diffusion process progressively reduces the distributional discrepancy between degraded depth prompts and their high-quality counterparts through Gaussian smoothing contraction, eventually converging both distributions to a shared isotropic Gaussian prior (see the 2D Gaussian analogy in Fig. 2). Since increasing noise also erases the prompt content, there exists an optimal timestep that balances prompt preservation and alignment with the high-quality depth distribution.

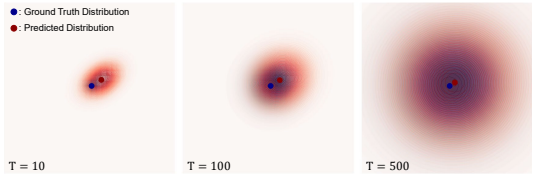


Figure 2: The distribution of the degraded depth prompt (with predicted mean and variance from coarse refinement) progressively aligns with the ground-truth distribution due to Gaussian smoothing contraction. Dots indicate the means.

Based on this insight, AdaDS operates in two stages. The first stage performs coarse prompt refinement to produce an initial depth estimate and a variance map that quantifies refinement uncertainty. This uncertainty is then used to adaptively select a timestep that preserves reliable prompt information while improving alignment with the high-quality depth prior. In the second stage, a dedicated noise sampling module predicts timestep-conditioned noise that, when added to the coarse estimate, places the intermediate latent in a high-density region of the high-quality depth distribution. This sample is then denoised by a pre-trained depth diffusion model to produce the final high-resolution metric depth map. By design, imperfect upstream prompt refinement can be tolerated, as adequate noising projects the sample onto the learned generative prior, yielding robustness to diverse depth degradations.

The main contributions of this work are:

- We introduce AdaDS, a generalizable framework for prompted metric depth estimation that effectively addresses robustness challenges under arbitrary upsampling factors, diverse degradation patterns, and unseen real-world scenes.
- We propose a novel adaptive sampling strategy that exploits the contraction property of the forward diffusion process to place an intermediate noisy sample in a high-density region of the high-quality depth distribution, thereby leveraging the powerful generative prior of a relative depth diffusion model for robust metric depth estimation.
- We conduct extensive experiments on both real-world and synthetic benchmarks, demonstrating AdaDS's superior zero-shot generalization and robustness compared to existing approaches.

2 Related Work

Depth Super-Resolution. Metric depth estimation from low-resolution depth prompts has commonly been studied as depth super-resolution (DSR). Early DSR works restore high-frequency details lost in synthetically downsampled depth maps [6, 52, 47]. The subsequent availability of

real-world datasets [8, 41] has motivated specialized designs for more complex degradations, such as structure flow guidance [46], degradation-driven depth bins [43], degradation-aware representation learning [37], and semantic-driven contrastive learning [45], along with other advanced architectures [36, 12]. Despite impressive in-domain performance, these methods are tailored to specific degradation assumptions and generalize poorly to unseen degradations and scenes.

Prompted Metric Depth Estimation. Recent works leverage depth foundation models to improve generalization across scenes and prompt patterns. PromptDA [18] integrates low-cost LiDAR prompts into the decoder of Depth Anything [44] at multiple scales. PriorDA [38] pre-fills diverse metric priors through pixel-level alignment with monocular predictions and refines them with a conditioned depth estimator. OMNI-DC [54] learns robust depth completion from synthetic data with diverse sparsity patterns. Despite strong zero-shot performance, these methods produce depth through deterministic feed-forward regression, so their robustness to prompt errors largely depends on the degradations simulated during training. Marigold-DC [32] instead injects sparse depth into a pre-trained depth diffusion model as test-time guidance via iterative optimization, which also enforces consistency with erroneous measurements. In contrast, AdaDS determines how much to trust the prompt via estimated uncertainty and projects its estimate onto a high-density region of the high-quality depth distribution, enabling robust metric depth estimation from arbitrarily degraded low-resolution depth prompts.

Depth Diffusion Models. Diffusion models have recently emerged as a powerful approach for high-fidelity depth estimation. Marigold [13] pioneered this direction by fine-tuning a Stable Diffusion [25] model on limited synthetic data, enabling strong zero-shot performance with fine-grained outputs. Subsequently, GeoWizard [4] jointly estimates depth and surface normals for enhanced geometric consistency; Lotus [7] mitigates undesirable variance in the generative process by directly predicting deterministic annotations; DepthFM [5] accelerates sampling through flow matching; and Jasmine [33] extends diffusion-based depth estimation to self-supervised learning. Furthermore, BetterDepth [51] and SharpDepth [22] combine the generalization of discriminative pre-training with the detail recovery capability of diffusion models, balancing accuracy and visual sharpness. Unlike these methods that generate depth from pure noise, AdaDS starts denoising from an adaptively selected intermediate timestep anchored to the refined prompt.

3 Depth Diffusion Model Preliminary

Our framework employs the pre-trained Marigold-LCM [13, 14] as the denoiser, benefiting from its strong modeling of high-quality depth distributions and efficient one-step inference capability. It is distilled from a Denoising Diffusion Probabilistic Model (DDPM) [9] based teacher model via latent consistency distillation [31, 21], and thus shares the same forward process as the teacher.

Specifically, given an RGB image $\mathbf{x}$ and its corresponding depth map $\mathbf{d}$, both are first encoded into the latent space $\mathbf{z}_x = \mathcal{E}_l(\mathbf{x})$ and $\mathbf{z}_d = \mathcal{E}_l(\mathbf{d})$ via a frozen VAE [16] encoder $\mathcal{E}_l(\cdot)$, where the depth map is affinely normalized to $[-1, 1]$ before encoding. In the forward diffusion process, the depth latent $\mathbf{z}_d$ is progressively corrupted by Gaussian noise according to the conditional distribution:

$$q(\mathbf{z}_t \mid \mathbf{z}_d) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\,\mathbf{z}_d,\ (1-\bar{\alpha}_t)\mathbf{I}\big), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s) \in (0, 1]$ is the cumulative product of the noise schedule β_s, and as $t \to T$, $\mathbf{z}_t$ approximately follows a standard Gaussian distribution. Consistency models [31] learn a consistency function f_θ that maps any noisy sample on a probability flow ODE trajectory directly to its origin. Accordingly, Marigold-LCM estimates the clean depth latent from a noisy latent at an arbitrary timestep t in a single step:

$$\hat{\mathbf{z}} = f_\theta\big(\mathbf{z}_x \| \mathbf{z}_t,\ t\big), \tag{2}$$

where $\|$ denotes channel-wise concatenation. Although standard inference starts from pure noise $\mathbf{z}_T = \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, this property allows denoising to be initiated from an intermediate timestep. The final depth map is then recovered as $\hat{\mathbf{d}} = \mathcal{D}_l(\hat{\mathbf{z}})$ via the pre-trained VAE decoder $\mathcal{D}_l(\cdot)$.

4 AdaDS Framework

AdaDS is motivated by the contraction property of the forward diffusion process. Let $\mathbf{d}_{in}$ denote the degraded low-resolution (LR) depth prompt and $\mathbf{z}_{in} = \mathcal{E}_l(\mathbf{d}_{in})$ its latent representation. As shown in

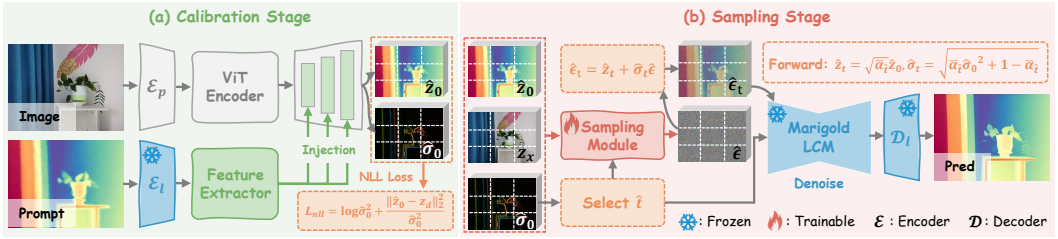


Figure 3: **Overall framework of AdaDS.** (a) The first stage produces a refined depth latent along with an estimated uncertainty map reflecting the refinement reliability. (b) The second stage blends the coarse refinement with estimated noise, placing it in a high-density region of the high-quality depth distribution. The resulting intermediate noisy latent is denoised by a pre-trained relative depth diffusion model to recover the final depth map. $\mathcal{E}_p$, $\mathcal{E}_l$ and $\mathcal{D}_l$ denote the patch embedding, latent encoder and latent decoder, respectively.

Eq. (1), the discrepancy between the forward distributions $q(\mathbf{z}_t \mid \mathbf{z}_d)$ and $q(\mathbf{z}_t \mid \mathbf{z}_{in})$ diminishes as $\bar{\alpha}_t$ approaches 0, while the prompt content is increasingly erased. This implies an optimal timestep that balances prompt fidelity and alignment with the high-quality depth distribution. Accordingly, AdaDS perturbs a refined depth latent with learned noise at the selected timestep, placing the noisy sample in a high-density region of the high-quality depth distribution learned by a pre-trained diffusion model, from which the final depth map is recovered by denoising. As illustrated in Fig. 3, AdaDS is implemented as a two-stage pipeline.

Calibration Stage. The calibration stage is illustrated in Fig. 3 (a). It refines the LR depth prompt with two objectives: (1) reducing the gap between the LR prompt and its high-quality counterpart, thereby avoiding excessively large timesteps that would erase important prompt content; and (2) estimating the predictive uncertainty of the refinement, which serves as the basis for adaptive timestep selection in the second stage.

To achieve these goals, we design a two-branch module $\mathcal{R}(\cdot)$ that takes the image $\mathbf{x}$ and interpolated $\mathbf{d}_{in}$ as inputs, producing a refined depth latent and its associated uncertainty: $(\hat{\mathbf{z}}_0, \hat{\sigma}_0) = \mathcal{R}(\mathbf{x}, \mathbf{d}_{in})$, where $\hat{\mathbf{z}}_0$ is the refined depth latent and $\hat{\sigma}_0$ denotes the predicted standard deviation at each spatial location. The image branch fine-tunes a variant of Depth Anything V2 [44] to exploit its strong spatial perception capability. It consists of a patch embedding layer $\mathcal{E}_p(\cdot)$ that downsamples the input image $\mathbf{x}$ by a factor of 16, followed by a sequence of ViT [3] blocks for global feature extraction, and a DPT [23] head for progressive feature upsampling. The depth branch first encodes $\mathbf{d}_{in}$ into the latent space $\mathbf{z}_{in} = \mathcal{E}_l(\mathbf{d}_{in})$ at $1/8$ scale. A UNet-like architecture [26] then extracts multi-scale features at $1/32$, $1/16$, and $1/8$ scales, which are injected into the DPT head of the image branch via element-wise addition to enhance depth-aware guidance. Training uses two complementary losses. First, we adopt the Negative Log-Likelihood (NLL) loss [15] for well-calibrated uncertainty estimation:

$$\mathcal{L}_{\text{nll}} = \log \hat{\sigma}_0^2 + \frac{\|\hat{\mathbf{z}}_0 - \mathbf{z}_d\|_2^2}{\hat{\sigma}_0^2}, \tag{3}$$

where $\mathbf{z}_d$ is the ground-truth depth latent. Second, to compensate for potential reconstruction errors introduced by the VAE encoder/decoder pair, we add an ℓ_1 supervision on the decoded depth:

$$\mathcal{L}_d = \|\mathcal{D}_l(\hat{\mathbf{z}}_0) - \mathbf{d}\|_1. \tag{4}$$

Sampling Stage. As shown in Fig. 3 (b), the second stage builds upon the outputs of the calibration stage to define an approximate forward diffusion process centered on the refined latent:

$$\hat{q}(\mathbf{z}_t \mid \hat{\mathbf{z}}_0, \hat{\sigma}_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\,\hat{\mathbf{z}}_0,\ \bar{\alpha}_t\hat{\sigma}_0^2 + 1 - \bar{\alpha}_t\big), \tag{5}$$

where the variance is defined element-wise. Before detailing the sampling procedure, we first introduce and prove a *proposition* that underpins our adaptive timestep selection strategy. Let $\mathcal{W}(\cdot,\cdot)$ denote the 2-Wasserstein distance, $\hat{q}_t$ the approximate forward distribution in Eq. (5), and q_t the ground-truth forward distribution in Eq. (1). As the forward process acts independently on each latent element, we analyze a single element in the following. For analytical tractability, we measure their

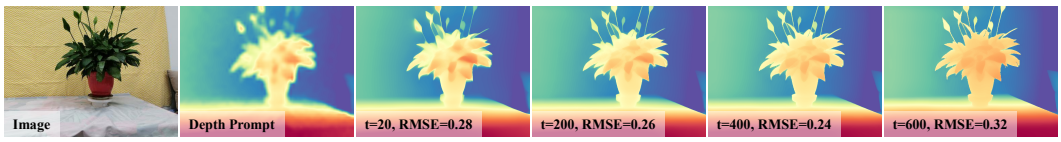


Figure 4: As t increases, the reconstructed depth maps exhibit sharper structural details but deviate further from the input depth prompt, so that the error first decreases and then increases. All depth maps share a unified color palette, with RMSE reported in centimeters.

misalignment with an upper bound $\overline{\mathcal{W}}_t \geq \mathcal{W}(\hat{q}_t, q_t)$ derived below, and define a simple surrogate objective with a closed-form maximizer to trade off distributional alignment and content preservation:

$$\mathcal{H}(\bar{\alpha}_t) = \sqrt{\bar{\alpha}_t} \cdot \exp\left(-\lambda \overline{\mathcal{W}}_t\right), \tag{6}$$

where the signal coefficient $\sqrt{\bar{\alpha}_t}$ in Eq. (1) measures content preservation and $\lambda > 0$ balances them.

Proposition 4.1. *For $\bar{\alpha}_t \in (0, 1]$, the function $\mathcal{H}(\bar{\alpha}_t)$ admits a unique global maximum.*

Proof. Coupling $\hat{q}_t$ and q_t through the same forward noise ϵ yields samples $\mathbf{z}'_t \sim \hat{q}_t$ and $\mathbf{z}_t \sim q_t$ with $\mathbf{z}'_t - \mathbf{z}_t = \sqrt{\bar{\alpha}_t}(\mathbf{z}_0 - \mathbf{z}_d)$, where $\mathbf{z}_0 \sim \mathcal{N}(\hat{\mathbf{z}}_0, \hat{\sigma}_0^2)$. Since the 2-Wasserstein distance is the infimum over all couplings, it is upper bounded by

$$\mathcal{W}(\hat{q}_t, q_t) \leq \sqrt{\mathbb{E}(\mathbf{z}'_t - \mathbf{z}_t)^2} = \sqrt{\bar{\alpha}_t}\, \omega \triangleq \overline{\mathcal{W}}_t, \tag{7}$$

where $\omega = \sqrt{(\hat{\mathbf{z}}_0 - \mathbf{z}_d)^2 + \hat{\sigma}_0^2}$. Substituting $\overline{\mathcal{W}}_t$ into Eq. (6) gives $\mathcal{H}(\bar{\alpha}_t) = \sqrt{\bar{\alpha}_t}\, e^{-\lambda\omega\sqrt{\bar{\alpha}_t}}$, whose derivative is

$$\frac{d\mathcal{H}}{d\bar{\alpha}_t} = \frac{e^{-\lambda\omega\sqrt{\bar{\alpha}_t}}}{2\sqrt{\bar{\alpha}_t}}\left(1 - \lambda\omega\sqrt{\bar{\alpha}_t}\right) \tag{8}$$

vanishes at the critical point $\bar{\alpha}_t^* = 1/(\lambda\omega)^2$. For $\bar{\alpha}_t < \bar{\alpha}_t^*$, the derivative is positive; for $\bar{\alpha}_t > \bar{\alpha}_t^*$, it is negative. Therefore, $\mathcal{H}(\bar{\alpha}_t)$ has a unique global maximum on $(0, 1]$ at $\min\{1, \bar{\alpha}_t^*\}$. □

This proposition shows that the trade-off between prompt fidelity and distributional alignment admits a unique optimal timestep, as illustrated in Fig. 4. During inference, since $\mathbf{z}_d$ is unavailable, we estimate ω from the predicted uncertainty. For well-calibrated uncertainty, $\mathbb{E}(\hat{\mathbf{z}}_0 - \mathbf{z}_d)^2 \approx \hat{\sigma}_0^2$, so that $\omega \propto \hat{\sigma}_0$. As a single timestep is shared by the entire latent, we average the predicted standard deviations over all N latent elements:

$$\bar{\sigma}_0 = \frac{1}{N}\sum_{i=1}^{N} \hat{\sigma}_0^{(i)}. \tag{9}$$

Absorbing the constant factor into $\tau \propto 1/\lambda$, the optimum becomes $(\tau/\bar{\sigma}_0)^2$. In practice, however, this rule drives $\hat{t}$ toward T and erases the prompt content too quickly. We therefore adopt its square-root relaxation, which decays more gently:

$$\bar{\alpha}_{\hat{t}} = \frac{\tau}{\bar{\sigma}_0}, \tag{10}$$

which is clamped to $[\bar{\alpha}_{min}, 1]$, consistent with the boundary case of Proposition 4.1. Given $\hat{t}$, a lightweight noise sampling network $\mathcal{S}(\cdot)$ takes as input the RGB latent $\mathbf{z}_x$, the mean $\hat{\mathbf{z}}_t = \sqrt{\bar{\alpha}_{\hat{t}}}\,\hat{\mathbf{z}}_0$ and standard deviation $\hat{\sigma}_t = \sqrt{\bar{\alpha}_{\hat{t}}\,\hat{\sigma}_0^2 + 1 - \bar{\alpha}_{\hat{t}}}$ of $\hat{q}_{\hat{t}}$, and outputs a noise prediction $\hat{\epsilon} = \mathcal{S}(\mathbf{z}_x, \hat{\mathbf{z}}_t, \hat{\sigma}_t)$. The intermediate noisy latent is formulated as

$$\hat{\epsilon}_t = \hat{\mathbf{z}}_t + \hat{\sigma}_t\, \hat{\epsilon}. \tag{11}$$

Note that although $\hat{t}$ is determined by the spatially averaged uncertainty $\bar{\sigma}_0$, the pixel-wise uncertainty $\hat{\sigma}_0$ is retained in $\hat{\sigma}_t$, so that less reliable regions receive stronger noise perturbation and are thus more strongly corrected by the diffusion prior. Finally, $\hat{\epsilon}_t$ is fed into the pre-trained Marigold-LCM model for one-step denoising at timestep $\hat{t}$ following Eq. (2), yielding the final depth latent $\hat{\mathbf{z}}_d = f_\theta(\mathbf{z}_x \| \hat{\epsilon}_t, \hat{t})$, which is decoded to the final high-resolution depth map $\hat{\mathbf{d}} = \mathcal{D}_l(\hat{\mathbf{z}}_d)$. To train $\mathcal{S}(\cdot)$ with Marigold-LCM kept frozen, we employ a reconstruction loss that encourages accurate recovery of both the latent and pixel-space depth:

$$\mathcal{L}_{\text{rec}} = \|\hat{\mathbf{z}}_d - \mathbf{z}_d\|_2^2 + \|\hat{\mathbf{d}} - \mathbf{d}\|_1. \tag{12}$$

Table 1: Quantitative zero-shot comparison on real-world benchmarks. Input resolutions for the low-resolution depth prompt are provided in the second row. For each metric, the best and second-best results are highlighted in **bold** and underline, respectively.

Method	Ref.	RGB-D-D						TOFDSR					
		192×144		64×48		32×24		192×144		64×48		32×24	
		RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
DA v2-L	NeurIPS-24	0.1082	0.0745	0.1082	0.0745	0.1082	0.0745	0.1035	0.0690	0.1035	0.0690	0.1035	0.0690
MG-LCM	CVPR-24	0.1306	0.1007	0.1306	0.1007	0.1306	0.1007	0.1219	0.0835	0.1219	0.0835	0.1219	0.0835
DA v3-NGL	ICLR-26	0.2835	0.2528	0.2835	0.2528	0.2835	0.2528	0.5229	0.3806	0.5229	0.3806	0.5229	0.3806
SGNet	AAAI-24	0.0726	0.0388	0.0723	0.0391	0.0740	0.0416	0.0726	0.0400	0.0726	0.0404	0.0767	0.0433
C2PD	AAAI-25	0.0724	0.0388	0.0717	0.0384	0.0714	0.0384	0.0723	0.0403	0.0718	0.0400	0.0722	0.0405
DuCos	ICCV-25	0.0744	0.0388	0.0741	0.0385	0.0898	0.0555	0.0742	0.0399	0.0742	0.0399	0.0958	0.0594
PromptDA	CVPR-25	0.0706	0.0506	0.0705	0.0505	0.0704	0.0503	0.0879	0.0623	0.0885	0.0624	0.0884	0.0620
OMNI-DC	ICCV-25	0.0680	0.0371	0.0687	0.0385	0.0846	0.0468	0.0748	0.0399	0.0890	0.0451	0.1071	0.0535
MG-DC	ICCV-25	0.0685	0.0382	0.0684	0.0392	0.0826	0.0475	0.0780	0.0416	0.0844	0.0441	0.0980	0.0508
PriorDA	ICLR-26	0.0655	0.0379	0.0613	0.0377	0.0691	0.0432	0.0731	0.0411	0.0755	0.0435	0.0833	0.0480
Ours	—	**0.0607**	**0.0332**	**0.0597**	**0.0330**	**0.0597**	**0.0342**	**0.0693**	**0.0336**	**0.0688**	**0.0346**	**0.0700**	**0.0363**

Additionally, we incorporate a gradient loss to encourage spatial smoothness and edge preservation in the predicted depth map:

$$\mathcal{L}_g = \|\partial_x \hat{\mathbf{d}} - \partial_x \mathbf{d}\|_1 + \|\partial_y \hat{\mathbf{d}} - \partial_y \mathbf{d}\|_1, \tag{13}$$

where ∂_x and ∂_y denote the horizontal and vertical spatial gradients, respectively.

4.1 Implementation Details

Dataset Details. We train AdaDS on the Hypersim dataset [24], which provides high-quality synthetic RGB-depth pairs rendered from photorealistic indoor scenes. After filtering out samples with large invalid depth regions, we obtain around 56K training samples. LR depth prompts are generated by downsampling the ground-truth depth with random scaling factors, followed by augmentations that mimic real-world sensor degradations, including Gaussian noise, random filtering, and sparse point removal, where the removed pixels are refilled by KNN-based inpainting.

Network Details. The calibration stage employs the ViT-Small variant of Depth Anything V2 as its image backbone, with the patch size changed to 16 to align its feature resolution with the latent space. Before encoding, depth maps are normalized to $[-1, 1]$ using the 2nd and 98th percentiles of valid depth values in the prompt, and the decoded prediction is mapped back to metric scale by the inverse transformation. Both the depth branch and the noise sampling module adopt lightweight UNet-like architectures built from ConvNeXt [20] blocks and transposed convolutions. Overall, AdaDS contains approximately 40.3M trainable parameters.

Training Details. We train the two stages of AdaDS independently, with the VAE and Marigold-LCM kept frozen. The calibration stage is trained for 42K iterations, and the sampling stage for 35K iterations, both with a batch size of 8. Training is performed on 4 NVIDIA RTX 4090 GPUs and takes approximately 14 hours in total. Following the OneCycle policy [30], the learning rate warms up from 10^{-5} to 10^{-4} over the first 10% of iterations and then decays to 10^{-5} with a cosine schedule. The two stages are trained with $\mathcal{L}_1 = \mathcal{L}_{\text{nll}} + \gamma \mathcal{L}_d$ and $\mathcal{L}_2 = \mathcal{L}_{\text{rec}} + \mathcal{L}_g$, respectively, where the balancing hyperparameter is set to $\gamma = 0.5$. For timestep selection, we set $\tau = 0.14$ and $\bar{\alpha}_{min} = 0.6$.

5 Experiment

5.1 Experiment Setup

We evaluate the zero-shot generalization of AdaDS on both real-world and synthetic benchmarks. For *real-world* evaluation, we use the official test splits of **RGB-D-D** [8] and **TOFDSR** [41], containing 405 and 560 samples, respectively. Both datasets pair LR depth maps captured by a mobile ToF camera at 192×144 with ground-truth depth acquired by a high-end ToF sensor at 512×384, and varying upsampling factors are simulated by further bicubic downsampling of the LR inputs.

Table 2: Quantitative zero-shot comparison on synthetic benchmarks. DS., GN., GB., SM. and LC. denote downsampling, Gaussian noise, Gaussian blur, sparse measurement and low-bit compression, respectively. RMSE metric is reported in this table.

Method	Ref.	ScanNet					NYUv2				
		16× DS.	DS. + GN.	DS. + GB.	DS. + SM.	DS. + LC.	16× DS.	DS. + GN.	DS. + GB.	DS. + SM.	DS. + LC.
DA v2-L	NeurIPS-24	0.1416	0.1416	0.1416	0.1416	0.1416	0.2075	0.2075	0.2075	0.2075	0.2075
MG-LCM	CVPR-24	0.1986	0.1986	0.1986	0.1986	0.1986	0.2587	0.2587	0.2587	0.2587	0.2587
DA v3-NGL	ICLR-26	0.2186	0.2186	0.2186	0.2186	0.2186	0.3095	0.3095	0.3095	0.3095	0.3095
PromptDA	CVPR-25	0.0927	0.0983	0.0935	0.0953	0.0950	0.1326	0.1352	0.1349	0.1369	0.1338
OMNI-DC	ICCV-25	0.1173	0.1224	0.1127	0.1188	0.1206	0.1966	0.2011	0.1894	0.2003	0.1980
MG-DC	ICCV-25	0.1222	0.1288	0.1173	0.1214	0.1262	0.2027	0.2077	0.1951	0.2025	0.2052
PriorDA	ICLR-26	0.0933	0.0949	0.0953	0.0954	0.0944	0.1432	0.1444	0.1474	0.1458	0.1439
Ours	—	**0.0754**	**0.0783**	**0.0810**	**0.0834**	**0.0773**	**0.1197**	**0.1217**	**0.1289**	**0.1315**	**0.1206**

For *synthetic* evaluation, we use **NYUv2** [29] (654 test samples) and **ScanNet** [2] (500 uniformly sampled test frames) at 640×480, where LR prompts are synthesized by bicubic downsampling followed by common sensor artifacts, including Gaussian noise, Gaussian blur, sparse measurement, and low-bit compression. Invalid ground-truth pixels are filled by colorization-based inpainting [17] before degradation, while all metrics are computed only on valid pixels. We report RMSE and MAE for absolute accuracy, and the threshold accuracy $\delta_{1.05}$, i.e., the percentage of pixels satisfying $\max(\hat{\mathbf{d}}/\mathbf{d}, \mathbf{d}/\hat{\mathbf{d}}) < 1.05$, for relative precision. For baselines, predictions of affine-invariant methods (DA v2-L and MG-LCM) are aligned to the ground truth via Theil–Sen regression, DA v3-NGL adopts its metric DA3NESTED-GIANT-LARGE-1.1 variant, and depth completion methods (OMNI-DC and MG-DC) take the LR depth values placed at their corresponding positions on the high-resolution grid as sparse inputs. Image-only methods thus yield identical results across prompt settings.

5.2 Comparison with the State-of-the-Art

We compare AdaDS against a range of state-of-the-art (SOTA) methods, including monocular depth estimation methods, i.e., Depth Anything V2 (DA v2-L) [44], Marigold-LCM (MG-LCM) [13] and Depth Anything V3 (DA v3-NGL) [19], DSR methods, i.e., SGNet [36], C2PD [12] and DuCos [42], as well as prompted depth estimation methods, i.e., PromptDA [18], OMNI-DC [54], Marigold-DC (MG-DC) [32] and PriorDA [38].

Real-World Evaluation. Quantitative results on the real-world benchmarks are reported in Tab. 1. AdaDS consistently outperforms all competing methods across every metric and upsampling factor. For instance, at the original input resolution, AdaDS outperforms the second-best methods, PriorDA on RGB-D-D and C2PD on TOFDSR, by 7.33% and 4.15% in RMSE, respectively. Notably, AdaDS exhibits remarkable stability with respect to input resolution changes, maintaining nearly consistent performance across all tested upsampling factors. In contrast, competing methods show clear brittleness. For example, the RMSE of MG-DC deteriorates by 20.58% when the upsampling factor increases from 2.7× to 16×. We further evaluate generalization to arbitrary (non-integer) upsampling factors in Fig. 5, where AdaDS achieves the highest $\delta_{1.05}$ under all factors.

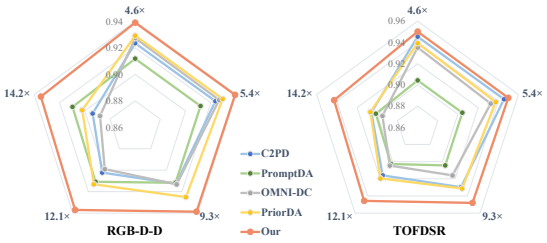


Figure 5: Comparison on real-world benchmarks with arbitrary upsampling factors. The $\delta_{1.05} \uparrow$ metric is reported in this figure.

Synthetic Evaluation. Quantitative results on synthetic benchmarks are summarized in Tab. 2, where the ground-truth depth is downsampled by 16× and further corrupted by different degradations. AdaDS significantly outperforms all baselines across all degradation types and both datasets. On average, it improves over the second-best methods by 16.46% (PriorDA) on ScanNet and 7.57% (PromptDA) on NYUv2, highlighting its strong robustness to compound depth degradations. Compared with the strongest image-only method, DA v2-L, AdaDS reduces RMSE by more than 36% in all settings, showing that metric cues remain valuable even when the depth prompts are degraded.

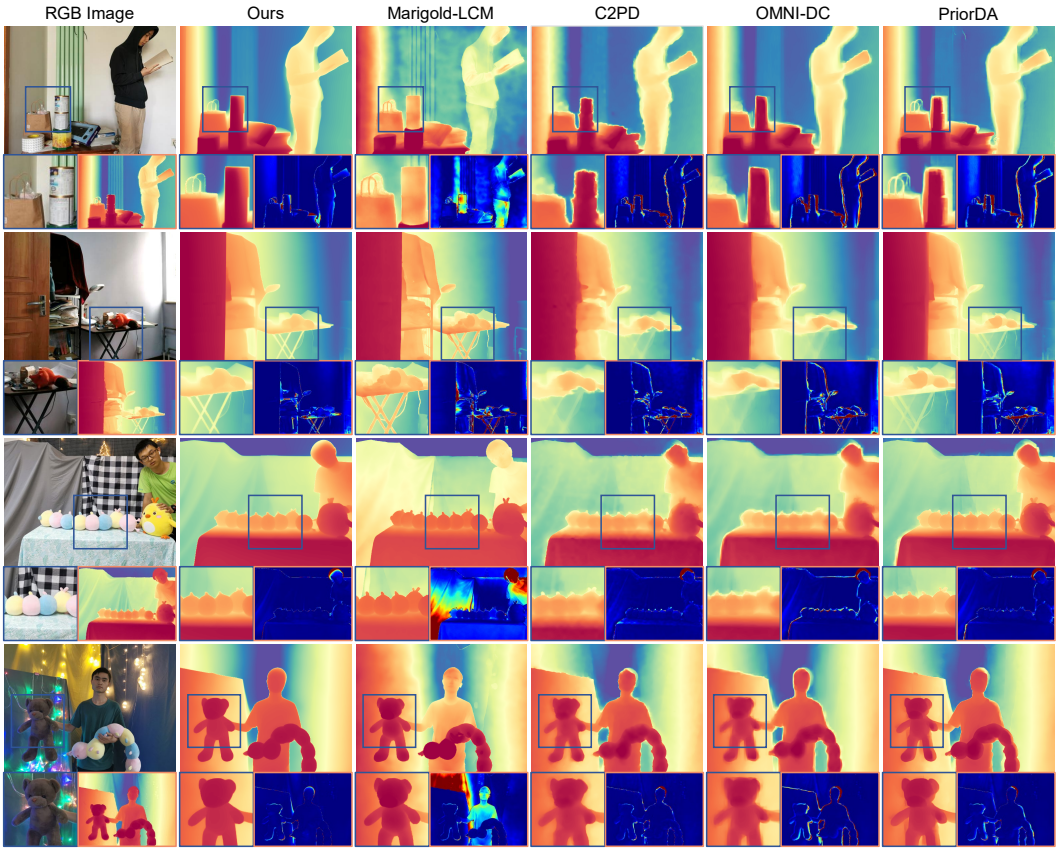


Figure 6: Qualitative comparison on real-world benchmarks. The first two samples are from RGB-D-D and the last two from TOFDSR. Blue boxes mark regions enlarged in the bottom-left inset of each result, and error maps are shown in the bottom-right inset, with warmer colors denoting larger errors.

Qualitative Comparison. Qualitative results on real-world benchmarks are presented in Fig. 6. AdaDS produces depth maps with substantially lower overall prediction error (as visualized in the error maps) while recovering fine-grained geometric details and sharp boundaries in the highlighted regions, outperforming all competitors. On synthetic benchmarks shown in Fig. 7, we back-project the predicted depth maps into 3D point clouds to better illustrate structural fidelity. Our method effectively restores severely distorted structures from heavily degraded inputs, preserving both local details and global scene geometry with noticeably higher quality than existing approaches.

6 Ablation Study

We perform comprehensive ablation studies on real-world benchmarks, with quantitative results reported in Tab. 3. Our calibration stage (b), built upon the DA v2-Small backbone (a), effectively fuses the LR depth prompt with RGB guidance and substantially improves over (a), reducing RMSE by 50.43% on RGB-D-D and 44.57% on TOFDSR. This confirms the importance of incorporating metric depth cues. To examine the role of diffusion sampling, variant (c) replaces the noise sampling and diffusion operations with direct regression of the final depth latent $\mathbf{z}_d$, leading to RMSE increases of 2.68% on RGB-D-D and 3.20% on TOFDSR compared with full AdaDS. Variant (d) further projects the predicted latent from (c) into the forward diffusion process and applies MG-LCM denoising without learned noise adaptation, yielding inferior results and showing that naive forward projection is insufficient. Replacing our adaptive timestep selection rule in Eq. (10) with uniformly random timesteps (e) causes a clear performance drop, validating the need to balance prompt preservation and distributional alignment. Similarly, substituting the learned noise prediction $\hat{\epsilon}$ with standard Gaussian noise (f) substantially degrades performance, demonstrating the effectiveness of the sampling module $\mathcal{S}(\cdot)$ in positioning the intermediate latent within the high-probability region of the target posterior.



Figure 7: Qualitative comparison on synthetic benchmarks. The first two samples are from NYUv2 and the last two from ScanNet. To facilitate 3D structural evaluation, the input and estimated depth maps are back-projected into point clouds and colorized with the corresponding RGB images.

	RGB-D-D		TOFDSR	
	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
(a) DA v2-S Baseline	0.1265	0.0900	0.1252	0.0740
(b) Our Calibration Stage	0.0627	0.0358	0.0694	0.0379
(c) (b) + Directly Regressing $\mathbf{z}_d$	0.0613	0.0346	0.0710	0.0379
(d) (c) + MG-LCM Denoising	0.0622	0.0367	0.0723	0.0391
(e) Replace $\hat{t}$ with Random t	0.0768	0.0497	0.0809	0.0455
(f) Replace $\hat{\epsilon}$ with $\mathcal{N}(0, \mathbf{I})$	0.0668	0.0387	0.0742	0.0394
Our	**0.0597**	**0.0330**	**0.0688**	**0.0346**

Table 3: Ablation study at $8\times$ factor. $\mathbf{z}_d$, $\hat{t}$ and $\hat{\epsilon}$ represent GT depth latent, predicted denoising timestep and sampling noise, respectively.

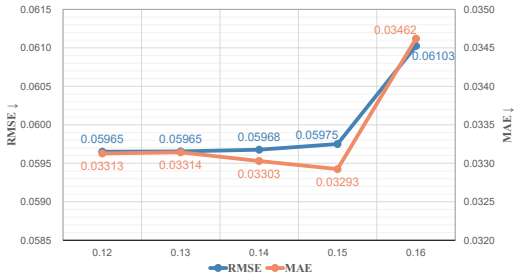


Figure 8: Validation experiment on timestep selection parameter τ. The x axis indicates τ values. All models are re-trained using indicated τ.

We further investigate the sensitivity to the threshold τ in the timestep selection rule. Results are shown in Fig. 8. AdaDS remains robust for small values of τ, but performance deteriorates significantly at $\tau = 0.16$. Larger τ selects very small timesteps that preserve input content but severely limit the model's ability to correct upstream errors. We adopt $\tau = 0.14$ in all experiments, as it provides the best trade-off between RMSE and MAE.

As MG-LCM supports multi-step sampling, we evaluate AdaDS with 1, 2, and 4 inference steps, as reported in Tab. 4. Our standard one-step setting achieves the best RMSE on both datasets. Although 2 steps slightly improve MAE on RGB-D-D and 4 steps achieve the best MAE on TOFDSR, both multi-step variants degrade RMSE. This suggests that additional denoising steps may enhance local detail but reduce metric fidelity. Since each step depends on the sampling network $\mathcal{S}(\cdot)$ to predict the intermediate sampling point, errors can accumulate over multiple steps and cause the prediction to drift from the true metric depth. Therefore, one-step denoising offers the best overall trade-off.

Fig. 9 further compares the intermediate output of the calibration stage with the final prediction. When the input depth is heavily degraded, the calibration stage recovers a large portion of the structural content and substantially narrows the distributional gap to high-quality depth, although residual inaccuracies remain. Based on this intermediate estimate, our adaptive noise sampling and diffusion denoising effectively correct these imperfections, producing a high-fidelity final depth map.

Inference Steps	RGB-D-D		TOFDSR	
	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
4 Steps	0.0657	0.0364	0.0733	**0.0331**
2 Steps	0.0605	**0.0329**	0.0698	0.0338
1 Step	**0.0597**	0.0330	**0.0688**	0.0346

Table 4: Ablation study of multi-step inference on real-world benchmarks.

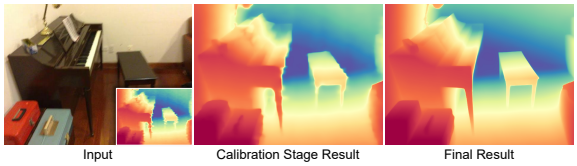


Figure 9: Visual comparison between the intermediate result from calibration stage and the final prediction.

η	κ	RGB-D-D		TOFDSR	
		RMSE ↓	MAE ↓	RMSE ↓	MAE ↓
0.1	0.5	0.0599	0.0333	0.0689	0.0353
0.2	0.25	0.0602	0.0338	0.0688	0.0355
0.3	0.125	0.0609	0.0346	0.0692	0.0360
Ours		0.0597	0.0330	0.0688	0.0346

Table 5: Robustness to perturbed calibration-stage estimates at $8\times$ factor. Larger η and smaller κ denote stronger perturbations.

Method	Time (s)	RGB-D-D	TOFDSR
MG-LCM [13]	0.199	0.1306	0.1219
ControlNet [50]	0.226	0.0664	0.0716
PriorDA [38]	0.108	0.0613	0.0755
Our Calibration Stage	0.015	0.0627	0.0694
Our Sampling Network	0.003	–	–
Ours Full	0.217	**0.0597**	**0.0688**

Table 6: Efficiency comparison at $8\times$ factor, reporting RMSE and per-image inference time on a single RTX 4090 GPU.

We further evaluate whether AdaDS can tolerate inaccurate calibration-stage estimates. Specifically, we perturb the refined latent as $\hat{\mathbf{z}}_0' = \hat{\mathbf{z}}_0 + \hat{\mathbf{z}}_0 \times \mathcal{N}(0, \eta)$ and suppress the predicted uncertainty as $\hat{\sigma}_0'^2 = \hat{\sigma}_0^2 \times \mathcal{U}(\kappa, 1)$, where larger η and smaller κ introduce stronger latent perturbation and more severe uncertainty underestimation, respectively. As shown in Tab. 5, AdaDS remains stable under all settings, with only minor degradation even under the strongest perturbation. This indicates that the sampling stage corrects imperfect calibration outputs rather than accumulating first-stage errors.

Finally, we compare computational efficiency in Tab. 6, where all diffusion-based methods use one-step denoising and ControlNet [50] denotes a Marigold-based variant conditioned on the upsampled depth prompt. AdaDS achieves the best accuracy with a modest overhead, as its calibration stage and sampling network take only 0.015s and 0.003s, respectively. The main cost comes from the MG-LCM denoiser, suggesting that AdaDS can benefit from faster diffusion backbones.

7 Limitations and Broader Impact

AdaDS is designed for low-resolution depth prompts and is trained and evaluated on indoor scenes, where such prompts are typically produced by ToF and structured-light sensors. It does not directly address monocular depth estimation or depth completion from sparse LiDAR or SfM measurements, whose uncertainty structures differ substantially. Moreover, as AdaDS relies on a pre-trained depth diffusion prior, its predictions may inherit biases or failure modes of this prior in rare scenes or domains far from its training distribution. Extending AdaDS to more general geometric guidance is a promising future direction. Regarding broader impact, AdaDS may reduce the reliance on expensive depth sensors and benefit robotics, AR/VR, and assistive perception. However, inaccurate estimates under out-of-distribution conditions could lead to unsafe decisions in safety-critical systems, and improved 3D perception may be misused for surveillance or privacy-invasive reconstruction, calling for careful validation and human oversight.

8 Conclusion

Generalization remains a key challenge in prompted metric depth estimation under diverse real-world degradations. We address this by exploiting the contraction property of the forward diffusion process, under which degraded depth prompts progressively align with their high-quality counterparts while the prompt content is gradually erased. Based on this trade-off, AdaDS calibrates the prompt with uncertainty estimation, adaptively selects the starting timestep, and injects learned noise to place the intermediate sample in a high-density region of the depth distribution learned by a pre-trained diffusion model. Extensive experiments demonstrate its strong zero-shot robustness to arbitrary upsampling factors, diverse degradations, and real-world depth prompts.

Acknowledgments and Disclosure of Funding

This work is supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (MOE-T2EP20124-0013), and partially supported by the Lee Kong Chian Fellowship awarded to ZHOU Pan by Singapore Management University.

References

[1] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *CVPR*, pages 811–820, 2024.

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *CVPR*, 2017.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[4] 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*, pages 241–258, 2024.

[5] 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*, pages 3203–3211, 2025.

[6] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing*, pages 2545–2557, 2018.

[7] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024.

[8] Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. In *CVPR*, pages 9229–9238, 2021.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, pages 6840–6851, 2020.

[10] Heng Jia, Linchao Zhu, and Na Zhao. H3r: hybrid multi-view correspondence for generalizable 3d reconstruction. In *ICCV*, pages 7655–7665, 2025.

[11] Shuyi Jiang, Qihao Zhao, Hossein Rahmani, De Wen Soh, Jun Liu, and Na Zhao. Gaussianblock: Building part-aware compositional and editable 3d scene by primitives and gaussians. In *ICLR*, pages 70188–70204, 2025.

[12] Jiahui Kang, Qing Cai, Runqing Tan, Yimei Liu, and Zhi Liu. C2pd: Continuity-constrained pixelwise deformation for guided depth super-resolution. In *AAAI*, pages 4212–4220, 2025.

[13] 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*, pages 9492–9502, 2024.

[14] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[17] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. *ACM Transactions on Graphics*, page 689–694, 2004.

[18] 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*, pages 17070–17080, 2025.

[19] 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.

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022.

[21] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[22] 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*, pages 17060–17069, 2025.

[23] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021.

[24] 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.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[27] Qiuhong Shen, Zike Wu, Xuanyu Yi, Pan Zhou, Hanwang Zhang, Shuicheng Yan, and Xinchao Wang. Gamba: Marry gaussian splatting with mamba for single-view 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[28] Wuxuan Shi, Mang Ye, and Bo Du. Symmetric uncertainty-aware feature transmission for depth super-resolution. In *ACM MM*, pages 3867–3876, 2022.

[29] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012.

[30] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386, 2019.

[31] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, pages 32211–32252, 2023.

[32] 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*, pages 5359–5370, 2025.

[33] Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *arXiv preprint arXiv:2503.15905*, 2025.

[34] Kun Wang, Zhiqiang Yan, Junkai Fan, Wanlu Zhu, Xiang Li, Jun Li, and Jian Yang. Dcdepth: Progressive monocular depth estimation in discrete cosine domain. *NeurIPS*, pages 64629–64648, 2024.

[35] Xinyi Wang, Xun Yang, Yanlong Xu, Yuchen Wu, Zhen Li, and Na Zhao. Affordbot: 3d fine-grained embodied reasoning via multimodal large language models. In *NeurIPS*, pages 140367–140387, 2026.

[36] Zhengxue Wang, Zhiqiang Yan, and Jian Yang. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *AAAI*, pages 5823–5831, 2024.

[37] 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*, pages 15813–15822, 2025.

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

[39] Yuchen Wu, Kun Wang, Yining Pan, and Na Zhao. Ccf: Complementary collaborative fusion for domain generalized multi-modal 3d object detection. In *CVPR*, pages 18745–18754, 2026.

[40] Jingqiao Xiu, Yicong Li, Na Zhao, Han Fang, Xiang Wang, and Angela Yao. Geometric alignment and prior modulation for view-guided point cloud completion on unseen categories. In *ICCV*, pages 27435–27444, 2025.

[41] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion. In *CVPR*, pages 4874–4884, 2024.

[42] 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*, pages 8361–8371. IEEE, 2025.

[43] Zhiqiang Yan, Zhengxue Wang, Jian Yang, and Gim Hee Lee. Degbins: Degradation-driven binning for depth super-resolution. *arXiv preprint arXiv:2605.09628*, 2026.

[44] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, pages 21875–21911, 2024.

[45] Xinchen Ye, Aokai Zhang, and Rui Xu. Semantics-driven contrastive learning for real-world depth super resolution. In *ACM MM*, pages 3085–3093, 2025.

[46] Jiayi Yuan, Haobo Jiang, Xiang Li, Jianjun Qian, Jun Li, and Jian Yang. Structure flow-guided network for real depth super-resolution. In *AAAI*, pages 3340–3348, 2023.

[47] Jiayi Yuan, Haobo Jiang, Xiang Li, Jianjun Qian, Jun Li, and Jian Yang. Recurrent structure attention guidance for depth super-resolution. In *AAAI*, pages 3331–3339, 2023.

[48] Jiayi Yuan, Haobo Jiang, De Wen Soh, and Na Zhao. Vggt-360: Geometry-consistent zero-shot panoramic depth estimation. In *CVPR*, pages 19874–19883, 2026.

[49] Zhihao Yuan, Shuyi Jiang, Chun-Mei Feng, Yaolun Zhang, Shuguang Cui, Zhen Li, and Na Zhao. Scene-r1: Video-grounded large language models for 3d scene reasoning without 3d annotations. *arXiv preprint arXiv:2506.17545*, 2025.

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

[51] 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*, pages 108674–108709, 2024.

[52] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pages 5697–5707, 2022.

[53] Yun Zhu, Jianjun Qian, Jian Yang, Jin Xie, and Na Zhao. Few-shot incremental 3d object detection in dynamic indoor environments. In *CVPR*, pages 18786–18795, 2026.

[54] Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. Omni-dc: Highly robust depth completion with multiresolution depth integration. In *ICCV*, pages 9287–9297, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the main contributions and scope of this paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations and future work in Section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We present a proposition in the Proposition 4.1 and provide a complete proof of it.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed implementation and experimental details in Sec. 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used in this paper are open access, and we plan to release our code at `https://github.com/w2kun/AdaDS`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed implementation and experimental details in Sec. 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports deterministic benchmark results following standard evaluation protocols, but does not include error bars or statistical significance tests due to the high computational cost of repeated training and evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information about the computer resources we used in our experiments in Sec. 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both potential positive societal impacts and negative societal impacts of our work in Sec. 7.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the creators of all external datasets, pre-trained models, and baseline methods used in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Our core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.