Edit2Perceive: Image Editing Diffusion Models Are Strong Dense Perceivers

Yiqing Shi[1*] Yiren Song[2*] Mike Zheng Shou[2†]
[1]Peking University, [2] National University of Singapore
* Equal contribution † Corresponding author.

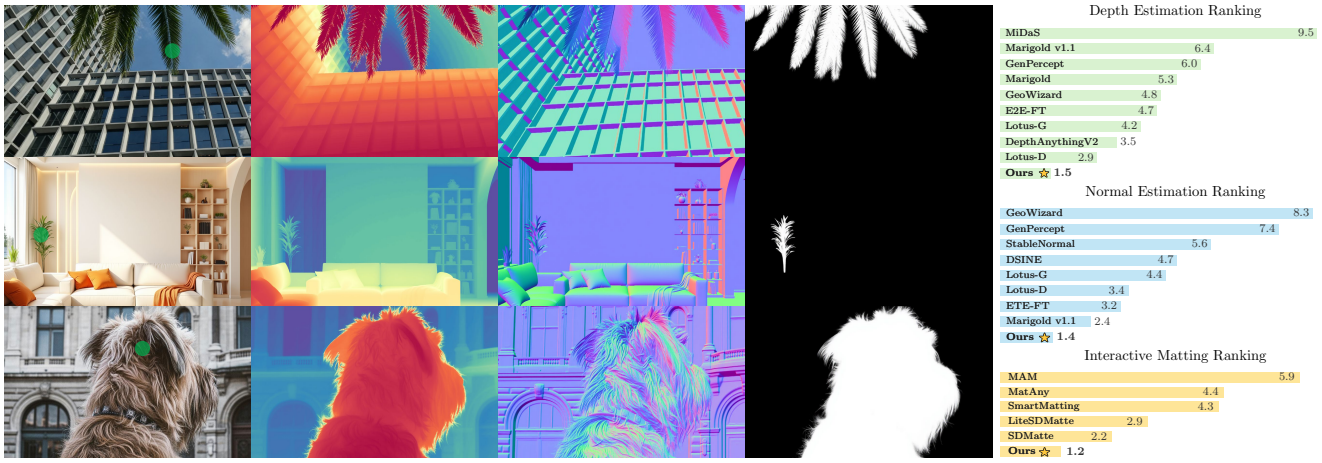


Figure 1. We present **Edit2Perceive**, a unified framework for diverse dense prediction tasks. Our model achieves state-of-the-art performance across zero-shot Monocular Depth Estimation, Surface Normal Estimation, and Interactive Matting, consistently outperforming previous methods. The bar charts on the right compare the average ranking score of our method against prior work on each task; lower is better. The green dot (●) on the input images indicates the positive user prompt for interactive matting.

Abstract

Recent advances in diffusion transformers have shown remarkable generalization in visual synthesis, yet most dense perception methods still rely on text-to-image (T2I) generators designed for stochastic generation. We revisit this paradigm and show that image editing diffusion models are inherently image-to-image consistent, providing a more suitable foundation for dense perception tasks. We introduce ***Edit2Perceive****, a unified diffusion framework that adapts editing models for depth, normal, and matting. Built upon the FLUX.1 Kontext architecture, our approach employs full-parameter fine-tuning and a pixel-space consistency loss to enforce structure-preserving refinement across intermediate denoising states. Furthermore, compared with conventional generative models, our single-step deterministic inference substantially reduces inference computational cost (FLOPs). Extensive experiments demonstrate comprehensive state-of-the-art results across all three tasks, revealing the strong potential of editing-oriented diffusion transformers for geometry-aware perception. Code is released at https://github.com/showlab/Edit2Perceive.*

1. Introduction

Dense perception tasks, such as Monocular Depth Estimation, Surface Normal Estimation, and Interactive Matting, are central to understanding 3D structure and object segmentation from 2D images. These tasks are particularly challenging because they require models to accurately predict pixel-level geometric or optical attributes. Monocular Depth Estimation, for instance, necessitates recovering 3D scene geometry from a single image, which is inherently ill-posed. Similarly, Surface Normal Estimation demands precise angular predictions, while Interactive Matting involves the segmenting foreground from the background in an image given interactive input, which is highly sensitive to edge details. Solving these ill-posed problems depends on how much prior knowledge a model possesses about the visual world. Recent progress has been largely driven by leveraging rich visual priors from large-scale text-to-image (T2I) diffusion models, such as Stable Diffusion [1], which have shown impressive success when adapted for perception tasks [2].

However, we argue that this paradigm is inherently limited by a representation mismatch. The T2I models are trained to synthesize diverse visual concepts based on un-

structured textual prompts, which excel in semantic composition ("concept-to-pixel" mapping) but lack the capacity to reason about structural relationships within an image. As a result, they are misaligned with the deterministic and geometry-aware nature of dense perception.

We revisit this assumption and propose that image-to-image (I2I) diffusion models, especially context-based editors such as FLUX.1 Kontext [3], offer a more natural foundation for dense perception. Unlike T2I models, these models are trained to generate semantically coherent edits conditioned on an existing image. This pretraining objective implicitly requires parsing the input into a structured scene representation—capturing objects, surfaces, and their interrelations. Such structured priors are essential for dense perception.

Building upon this insight, we introduce **Edit2Perceive**, a unified framework that adapts a powerful I2I diffusion model into a dense perceiver. We first convert the stochastic denoising process into a pseudo-deterministic path by fixing the random seed, ensuring a unique and reproducible input-output mapping. To further strengthen geometric fidelity, we design a pixel-space consistency loss that enforces fine-grained geometric fidelity and enhances structure preservation in the final output. Additionally, to meet the high numerical stability required for Monocular Depth Estimation task, we theoretically derive that a square-root mapping minimizes the Absolute Relative (AbsRel) Error caused by the preprocess. Finally, benefiting from the inherent mechanism of flow matching, Edit2Perceive enables single-step deterministic inference, combining accuracy with efficiency.

Extensive experiments across three major dense perception tasks (Monocular Depth Estimation, Surface Normal Estimation, and Interactive Matting) show that our method achieves state-of-the-art performance while using limited train data. It achieved impressive results in the wild (Fig. 1).

Our main contributions are summarized as follows:

- We demonstrate that image editing diffusion models, rather than text-to-image generators, provide a better inductive bias for deterministic dense perception, effectively bridging generative modeling and geometric reasoning.
- We propose **Edit2Perceive**, a unified diffusion transformer fine-tuned across depth, normal, and matting tasks, enhanced with a pixel-consistency loss and an analytically derived optimal normalization for depth.
- We achieve state-of-the-art performance with efficient single-step inference, showing that editing-oriented diffusion models can serve as a new class of perception-oriented foundation models.

2. Related Work

2.1. Image Generation and Editing Models

Large-scale diffusion models have become the cornerstone of visual content creation. Text-to-image (T2I) systems such as Stable Diffusion [1] and FLUX.1 [4] demonstrate strong semantic understanding through large-scale image–text pre-training. Meanwhile, research attention is shifting toward more controllable generation. Unlike T2I models that generate scenes from abstract text, I2I models aim to make precise modifications to existing visual content. Recent editing systems such as Step1X-Edit [5] and FLUX.1 Kontext [3] learn complex correspondences between image structure and editing instructions, achieving high structural consistency. With the transition to Diffusion Transformers [6], controllable generation has advanced rapidly. Models such as EasyControl [7] exemplify MM-DiT–based conditioning and have inspired a broad line of DiT-driven methods [8–17].

2.2. Dense Perception Tasks

Dense perception tasks, including monocular depth estimation, surface normal estimation, and interactive matting, are fundamental to computer vision. Traditional approaches can be broadly categorized into two paths. The first is data-driven, involving training specialized models on massive annotated datasets. From MiDaS [18], which pioneered cross-dataset generalization, to the Depth Anything series [19, 20], which leverages large-scale unlabeled data for pre-training, this path improves performance by scaling up data and model size. However, this approach is computationally intensive, and the resulting models are often designed for a single task, limiting their ability to generalize to new ones. The second path is task-driven, focusing on designing sophisticated inductive biases for specific tasks, such as the geometric constraints in DSINE [21] for normal estimation. While efficient, these designs are often difficult to transfer to other perception tasks. Both paths highlight the need for a powerful and general-purpose visual prior, which is precisely what large-scale generative models can offer.

2.3. Diffusion Models for Dense Perception

To leverage the powerful priors of generative models, an emerging line of research adapts diffusion models for dense perception tasks. Pioneering works like Marigold [2, 22] and GeoWizard [23] demonstrated the feasibility of fine-tuning pre-trained T2I models (e.g., Stable Diffusion) for depth and normal estimation, achieving excellent generalization with limited data. This paradigm was subsequently refined by a series of studies: GenPercept [24], Stable Normal [25], Lotus [26], and E2E-FT [27] improved the sampling process or training objectives to enable efficient single-step inference, significantly enhancing the models' practicality. Similarly, in the domain of image matting,

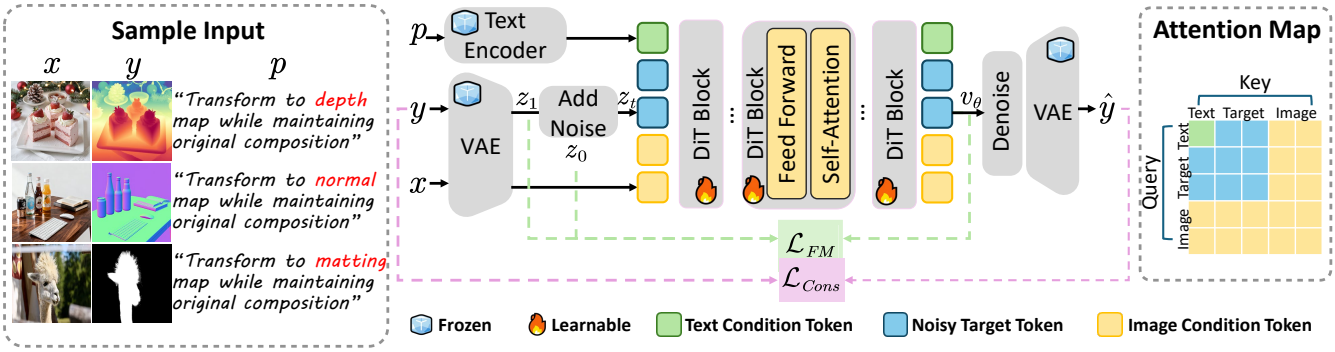


Figure 2. **Overview of the Edit2Perceive Framework.** We adapt the FLUX.1 Kontext editor for dense perception. An input image x and text prompt p condition for the target image y. In the forward process, target token will be mixed with noise, and get noisy target token z_t and then concatenate with text condition token and image condition token. The DiT backbone is trained to predict the velocity of the flow from a noise vector z_0 to the target latent z_1. Our training objective combines the latent-space flow matching loss ($\mathcal{L}_{\text{FM}}$) with a pixel-space consistency loss ($\mathcal{L}_{\text{Cons}}$) for enhanced geometric fidelity.

methods such as MAM [28], SMat [29], and SDMatte [30] have also successfully utilized diffusion priors to improve the modeling of fine-grained edge details.

While these works have collectively advanced the field, they are mostly built upon a common foundation: using T2I generative models as the starting point. In contrast, we explore an alternative possibility. Based on the "Representation Mismatch" argument from our introduction, we posit that I2I editing models, pre-trained for structured semantic understanding, can provide a more natural foundation for dense perception tasks. Concurrent to our work, FE2E [31] also builds on editing diffusion models, focusing on deterministic zero-noise denoising and consistent-velocity training, with logarithmic depth normalization to address precision issues. In contrast, our approach emphasizes geometric fidelity, introducing a pixel-space consistency loss and theoretically deriving a square-root depth mapping, while retaining Gaussian noise to better leverage the editor's generative priors.

3. Method

3.1. Edit2Perceive: Overall Architecture

Our framework, **Edit2Perceive**, is built upon FLUX.1 Kontext [3], a flow-matching [32] model based on the DiT [6] architecture. FLUX.1 Kontext unifies image generation and editing by incorporating semantic context from both text and image inputs through a simple sequence concatenation mechanism. It processes these inputs with a combination of shared and task-specific transformer blocks to extract image features, and generates new views that maintain semantic consistency, which is vital for dense perception tasks.

We formalize dense perception as a conditional diffusion editing task. As illustrated in Fig. 2, given an input RGB image $x \in \mathbb{R}^{H\times W\times 3}$ and a text prompt p, our goal is to predict the corresponding target image $y \in \mathbb{R}^{H\times W\times 3}$. Our method operates in the latent space of a pre-trained VAE. First, we encode the inputs: the conditioning image x is encoded into a latent representation c_x, the target dense map y into a target latent z_1, and the text prompt p into a text embedding c_p.

Next, we define a straight-line trajectory from a standard Gaussian noise vector $z_0 \sim \mathcal{N}(0, \mathbf{I})$ to the target latent z_1 using Rectified Flow [33]:

$$z_t = (1-t)z_0 + tz_1, \quad t \in [0,1]. \tag{1}$$

The constant velocity vector of this path is $\boldsymbol{v} = z_1 - z_0$. Our DiT model is trained to predict this velocity, denoted as $\boldsymbol{v}_\theta$ conditioned on the intermediate state z_t, the timestep t, and the conditions c_x and c_p

Training Objective. Our model is trained by optimizing a composite loss function that combines a latent-space flow matching with a pixel-space consistency loss. The primary objective is the flow matching loss, which minimizes the mean squared error between the predicted and true velocities:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,z_1,z_0,c} \left\| \boldsymbol{v}_\theta(\text{concat}(z_t, c_x, c_p), t) - \boldsymbol{v} \right\|_2^2. \tag{2}$$

While $\mathcal{L}_{\text{FM}}$ effectively guides the generation process in the latent space, it lacks direct supervision on the final pixel-level output, which can lead to geometric inaccuracies. To address this, we introduce a pixel-space consistency loss, $\mathcal{L}_{\text{Cons}}$, to enhance geometric fidelity. Our final training objective is a weighted sum of these two losses:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda \mathcal{L}_{\text{Cons}}, \tag{3}$$

Here, λ is an adaptive weight that balances the two objectives. The detailed formulation of $\mathcal{L}_{\text{Cons}}$ for different dense perception tasks and the curriculum-based strategy for weighting λ will be elaborated in Section 3.2.

Inference. During inference, we generate the target latent $\hat{z}_1$ by solving the ordinary differential equation (ODE) that defines the flow. Although Flow-Matching usually fails in one-step inference, geometric perception is a highly deterministic task, and we adapt a single-step Euler integrator:

$$\hat{z}_1 = z_0 + \boldsymbol{v}_\theta(\text{concat}(z_0, c_x, c_p), t = 0). \tag{4}$$

The final dense map $\hat{y}$ is then obtained by the VAE decoder. Furthermore, to enhance the determinism of the process, both training and inference share a fixed random seed for the initial noise z_0. Following Marigold [2], we also adopt an annealed multi-resolution noise strategy.

3.2. Enhancing Geometric Fidelity with Consistency Loss

Although $\mathcal{L}_{\text{FM}}$ aligns the flow in the latent space, it provides no direct supervision on the final pixel-level output reconstructed by the VAE decoder. Minor errors in the latent space can be amplified after decoding, leading to blurriness or structural artifacts.

To bridge this gap and enhance geometric fidelity, we introduce **pixel-space consistency loss**, $\mathcal{L}_{\text{Cons}}$, computed directly between the decoded prediction $\hat{y}$ and the ground truth y. This loss is customized for each task's specific properties. For simplicity, all formulations below are averaged over all pixels, denoted by $\mathbb{E}$.

Monocular Depth Estimation. We employ a Scale-and-Shift Invariant L1 Loss. We first align the prediction $\hat{y}$ to the ground truth y via least-squares fitting to obtain $\hat{y}_{\text{align}} = s\hat{y} + t$, and then compute the L1 error:

$$\mathcal{L}_{\text{Cons}}^{\text{depth}} = \mathbb{E}\left[|y - \hat{y}_{\text{align}}|\right]. \tag{5}$$

Surface Normal Estimation. To ensure accurate normal orientation while maintaining numerical stability, we use a mean angular error based on *atan2*. While the traditional *arccos* method suffers from gradient explosion when vectors are nearly collinear, our approach robustly computes the angle by using the dot and cross products to find its cosine and sine values, respectively:

$$\mathcal{L}_{\text{Cons}}^{\text{normal}} = \mathbb{E}\left[\text{atan2}\left(|y \times \hat{y}|, y \cdot \hat{y}\right)\right]. \tag{6}$$

This formulation is mathematically equivalent to *arccos* but gradient-stable.

Interactive Matting. Following standard practice [34], we compute separate L1 loss for the unknown transition region $\mathcal{U}$ and the known foreground / background regions $\mathcal{K}$, allowing the model to capture challenging edge details:

$$\mathcal{L}_{\text{Cons}}^{\text{matting}} = \mathbb{E}_{i\in\mathcal{U}}[|y_i - \hat{y}_i|] + \mathbb{E}_{i\in\mathcal{K}}[|y_i - \hat{y}_i|]. \tag{7}$$

Adaptive Loss Weighting. The consistency loss $\mathcal{L}_{\text{Cons}}$ provides a valuable geometric gradient from the pixel space to guide the latent-space flow matching. We combine the two losses with an adaptive weight λ:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda\mathcal{L}_{\text{Cons}}, \tag{8}$$

The weight λ is critical and is designed to implement a curriculum. It is set to zero during the first epoch to prioritize learning from the diffusion prior, and then linearly increases to balance the two objectives. This is formulated as:

$$\lambda = \frac{\text{sg}(|\mathcal{L}_{\text{FM}}|)}{\text{sg}(|\mathcal{L}_{\text{Cons}}|) + \epsilon} \cdot \max\left(0, \frac{\text{step}}{N_{\text{step}}} - 1\right), \tag{9}$$

where 'sg' denotes the stop-gradient operator and N_{step} is the total number of steps in one epoch, ϵ is a small number (0.001) added for numerical stability to prevent division by zero. This formulation enables the model focus on latent space flow matching at initial epoch and then gradually turns into a pixel space consistency loss.

3.3. Task-Specific Data Representation

To seamlessly adapt the supervision signal y of various dense perception tasks to the input requirements of the pre-trained editing model (a 3-channel BF16 tensor in the $[-1, 1]$ range), we design task-specific pre-processing pipelines aimed at maximizing information preservation and minimizing quantization error.

Monocular Depth Estimation. Depth maps (y) are single-channel and exhibit a long-tailed distribution. Direct linear normalization of y under BF16 precision causes significant quantization error in near-field details. To address this, we seek an optimal non-linear mapping, $g(y)$, that minimizes the quantization-induced relative error.

We formalize this by minimizing the integral of the relative error over the depth range $[y_{\min}, y_{\max}]$. Detailed in Appendix, this objective simplifies to minimizing the following integral with respect to the function g:

$$\min_g \frac{1}{512(y_{\max} - y_{\min})} \int_{y_{\min}}^{y_{\max}} \frac{g(y)_{\max} - g(y)_{\min}}{y \cdot g'(y)} \mathrm{d}y. \tag{10}$$

By applying the Cauchy-Schwarz inequality, we theoretically prove that this integral is minimized when $g'(y) \propto 1/\sqrt{y}$. This yields the optimal mapping function:

$$g(y) = \sqrt{y}. \tag{11}$$

Let $y_{\text{sqrt}} = g(y)$. We then apply a robust percentile-based linear normalization to these mapped values to scale them into the $[-1, 1]$ range for the VAE:

$$y_{\text{norm}} = \left(\frac{y_{\text{sqrt}} - y_{\text{sqrt, p2}}}{y_{\text{sqrt, p98}} - y_{\text{sqrt, p2}}} - 0.5\right) \times 2. \tag{12}$$

Here, $y_{\text{sqrt, p2}}$ and $y_{\text{sqrt, p98}}$ represent the 2nd and 98th percentiles of the sqrt-transformed image y_sqrt. Finally, the single-channel y_{norm} is replicated to three channels to form the final input representation for our framework.

Surface Normal Estimation. Normal maps are inherently 3-channel vectors. We simply ensure they are normalized to unit length: $y_{\text{norm}} = y/\|y\|_2$.

Interactive Matting. Alpha mattes are single-channel images in the $[0, 1]$ range. We first binarize the matte and then linearly map it to the $[-1, 1]$ range:

$$y_{\text{norm}} = (\mathbb{I}(y > 0.5) - 0.5) \times 2, \tag{13}$$

then replicating it to three channels.

4. Experiment

4.1. Experiment Setup

Implementation Details. Our experimental framework is built upon the FLUX.1 Kontext backbone. All parameters are frozen except DiT. For depth estimation and normal estimation tasks, we adopt the AdamW optimizer with a learning rate of 3×10^{-5} and a batch size of 16. For interactive matting, due to the higher memory demand from paired context images, we use the memory-efficient AdamW8bit optimizer with a batch size of 16. To accelerate convergence, we apply annealed multi-resolution noise during training. The model typically converges within approximately 6000 training steps. Each task is trained on a single NVIDIA H200 GPU for about 1.5 days.

Dataset and Benchmarks. We train and evaluate our model on task-specific benchmark datasets. **Monocular Depth Estimation:** Trained on Hypersim [35] and Virtual KITTI 2 [36], with a 90%:10% sample ratio. **Surface Normal Estimation:** Trained on Hypersim [35], InteriorVerse [37], and Sintel [38], with a 50%:45%:5% sample ratio. **Interactive Matting:** Trained on AM-2k [39], Distinctions-646 [40], Composition-1k [41], and COCO-Matting [42], with a 25%:25%:25%:25% sample ratio.

For evaluation, we test our models on zero-shot generalization on all benchmarks except for AM-2k in Interactive Matting task. **Monocular Depth Estimation:** Evaluated on NYUv2 [43], KITTI [44], ETH3D [45], ScanNet [46], and DIODE [47]. **Surface Normal Estimation:** Evaluated on NYUv2 [43], ScanNet [46], iBims-1 [48], and DIODE [47]. **Interactive Matting:** Evaluated on P3M-500-NP [49], AM-2k [39], and AIM-500 [50].

Metrics. For Monocular Depth Estimation, we report *AbsRel* and δ_1. Specifically, AbsRel $= \mathbb{E}[(\hat{y}_{align} - y)/y]$ where $\mathbb{E}$ is the expectation (average) of all valid pixels, $\hat{y}_{align}$ denotes the aligned prediction, and y the ground-truth depth. δ_1 measures the proportion of pixels satisfying $\max(\hat{y}_{align}/y, \quad y/\hat{y}_{align}) < 1.25$.

For Surface Normal Estimation, we report Mean angular error and the percentage of pixels with error ¡ 11.25°.

For Interactive Matting, we evaluate using MSE, MAD, SAD, Gradient, and Connectivity metrics.

4.2. Quantitative Evaluation

Table 1, 2, and 3 demonstrate the quantitative results across three dense perception tasks. Despite using far less training data and a single-step inference, our method consistently outperforms all strong baselines, demonstrating the effectiveness of **image-editing diffusion models** in structural understanding and geometric reasoning.

For Zero-Shot Monocular Depth Estimation, as shown in Table 1, our approach achieves new SOTA performance on five benchmarks, reaching an average rank of 1.5. On ETH3D and Scannet, the AbsRel error decreases 27% and 11% compared to second-best model, respectively. Notably, our model is trained on only 74K images yet surpasses heavily supervised approaches such as *DepthAnything V2* (62.6M images), highlighting the efficiency of our depth representation optimization.

For Zero-Shot Surface Normal Estimation (Table 2), our model ranks first on all benchmarks with an average rank of 1.4, confirming strong cross-dataset robustness and superior geometric detail reconstruction.

Finally, as shown in Table 3, our interactive variant achieves the lowest errors on all three datasets and an average rank of 1.2, demonstrating precise alpha prediction and excellent generalization to both portrait and non-portrait scenes.

4.3. Qualitative Evaluation

In this section, we present the qualitative evaluation results. As shown in Fig. 3, we compare our method with baseline approaches on the Monocular Depth Estimation, Surface Normal Estimation and Interactive Matting. Our model produces depth maps with finer structural details and stronger consistency with the input image, while baseline methods tend to miss or blur small but critical objects.

Similar improvements are observed in the surface normal and matting tasks. Our results exhibit better spatial coherence and visual consistency, demonstrating the model's capability to preserve both geometric accuracy and perceptual quality across dense perception tasks.

4.4. Ablation Study

To thoroughly analyze the contributions of each component in our framework, we conduct a series of detailed ablation studies. We first validate our core thesis on the superiority of I2I editing models as a foundation, then separately evaluate the effectiveness of our proposed pixel-space consistency loss and the theoretically optimal depth mapping function. Further details are provided in Supplementary Sections **B.1** and **C.1** .

Importance of the Base Model: Image-to-Image (I2I) vs. Text-to-Image (T2I). Our core thesis posits that due to differences in pre-training objectives, I2I models designed for editing are better suited for dense perception than T2I

Table 1. Quantitative comparison for zero-shot monocular depth estimation benchmarks. Our method is evaluated against recent SOTA methods across five benchmarks. The best and second-best are highlighted. All metrics are presented in percentage terms.

Method	Data	NYU		KITTI		ETH3D		Scannet		DIODE		AvgRank↓
		AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	
MiDaS	2M	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5	9.5
GeoWizard	280k	5.2	96.6	9.7	92.1	6.4	96.1	6.1	95.3	29.7	79.2	4.8
GenPercept	74k	5.6	96.0	9.9	90.4	6.2	95.8	-	-	35.7	75.6	6.0
DepthAnything V2	62.6M	4.5	97.9	7.4	94.6	13.1	86.5	-	-	26.5	73.4	3.5
Marigold	74k	5.5	96.4	9.9	91.6	6.5	95.9	6.4	95.2	30.8	77.3	5.3
Marigold1.1	74k	5.5	96.4	10.5	90.2	6.9	95.7	5.8	96.3	29.8	78.2	6.4
Lotus-D	59k	5.1	97.2	8.1	93.1	6.1	97.0	5.5	96.5	22.8	73.8	2.9
Lotus-G	59k	5.4	96.8	8.5	92.2	5.9	97.0	5.9	95.7	22.9	72.9	4.2
E2E-FT	74k	5.4	96.5	9.6	92.1	6.4	95.9	5.8	96.5	30.3	77.6	4.7
Edit2Perceive	74k	4.4	97.6	7.9	94.5	4.3	98.3	4.9	97.3	24.8	81.4	1.5

Table 2. Quantitative evaluation of our method on zero-shot surface normal estimation benchmarks. The best and second-best performances are highlighted.

Method	Data	NYU		Scannet		iBims-1		DIODE		AvgRank↓
		Mean↓	11.25°↑	Mean↓	11.25°↑	Mean↓	11.25°↑	Mean↓	11.25°↑	
GeoWizard	278k	19.0	50.0	17.6	54.6	19.3	62.3	24.7	30.1	8.3
GenPercept	44k	18.3	56.0	18.2	57.4	18.3	63.8	22.3	38.1	7.4
StableNormal	278k	17.8	54.2	16.7	54.0	17.1	67.8	19.3	53.8	5.5
DSINE	160K	16.2	61.0	16.4	59.6	17.1	67.4	19.9	41.8	4.5
Lotus-D	59k	16.2	59.8	14.7	64.0	17.1	66.4	-	-	3.1
Lotus-G	59k	16.5	59.4	15.1	63.9	17.2	66.2	-	-	4.4
ETE-FT	74k	16.5	60.4	14.7	66.1	16.1	69.7	19	44.4	3.0
Marigold v1.1	77k	16.1	60.5	14.5	66.1	16.3	68.5	18.8	45.5	2.4
Edit2Perceive	77k	15.7	61.6	14.1	66.3	15.1	70.9	18.7	44.3	1.4

Table 3. Quantitative evaluation of our method on interactive matting benchmarks. The best and second-best performances are highlighted. All methods use the same visual input points.

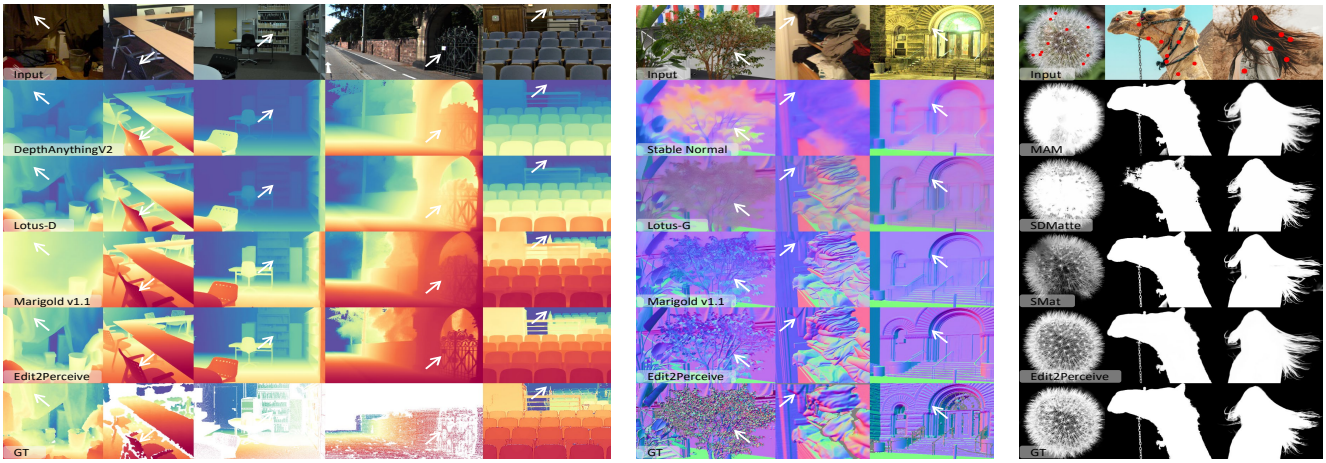
Method	AIM-500					P3M-500-NP					AM-2k					AvgRank↓
	MSE↓	MAD↓	SAD↓	Grad↓	Conn↓	MSE↓	MAD↓	SAD↓	Grad↓	Conn↓	MSE↓	MAD↓	SAD↓	Grad↓	Conn↓	
MAM	0.075	0.108	186.5	37.5	40.4	0.087	0.116	207.5	29.4	43.5	0.060	0.081	141.6	22.5	31.5	5.9
MatAny	0.043	0.052	87.0	33.4	25.4	0.029	0.034	57.3	25.9	16.0	0.012	0.019	32.2	15.7	20.4	4.4
SmartMatting	0.030	0.039	66.3	46.6	18.8	0.024	0.029	50.5	28.5	19.6	0.030	0.037	62.6	33.8	15.9	4.3
LiteSDMatte	0.011	0.021	34.4	24.3	20.0	0.012	0.017	29.9	16.6	21.8	0.009	0.016	27.5	13.6	17.7	2.9
SDMatte	0.011	0.019	31.8	26.8	17.5	0.013	0.018	32.0	20.4	20.8	0.006	0.010	17.5	13.2	10.9	2.2
Edit2Perceive	0.006	0.017	29.1	18.2	15.7	0.003	0.011	19.4	13.2	10.2	0.004	0.012	20.4	9.6	9.9	1.2

generators. To provide direct experimental evidence, we conduct a rigorous controlled experiment.

We select two models with identical architectures: **FLUX.1 Kontext (I2I)** and **FLUX.1 (T2I)**. Crucially, we construct the same I2I-style fine-tuning pipeline for the T2I model (FLUX.1) as our main framework, using token concatenation for the text prompt, condition image and target image. This ensures that the only significant difference between the two models is the prior knowledge acquired during their respective pre-training phases.

As shown in Table 4 and Table 5, we conducted a total of six sets of controlled experiments (four for depth estimation and two for normal estimation). The results are unequivocal: across all experimental settings, the performance of the I2I-based model comprehensively and significantly surpasses its T2I counterpart. For instance, in the depth estimation task (Table 4), even with the most basic configuration (ID 5 & 1), the I2I model achieves a remarkable **25%** and **27%** relative improvement in AbsRel on NYUv2 and KITTI, respectively. This advantage persists even after applying all our optimization strategies (ID 8 & 4). The same trend is observed in the normal estimation task (Table 5). This overwhelming experimental evidence strongly validates our core thesis: the structured semantic priors learned by I2I models provide a far superior starting point for downstream perception tasks than those from T2I models.

As shown in **Fig. 4**, we visualized **Self-Attention Maps** (16th DiT block) in . I2I capture sharp object boundaries as early as Epoch 1, while T2I remains scattered (Similar in Epoch 3). This suggests I2I possess a superior "structural prior" for geometry tasks.



(a) Monocular Depth Estimation (b) Surface Normal Estimation (c) Interactive Matting

Figure 3. Qualitative Comparison of our methods with other SOTA methods across different benchmarks. The arrows emphasize the regions that Edit2Perceive (ours) significantly outperform others. Zoom in for better view.

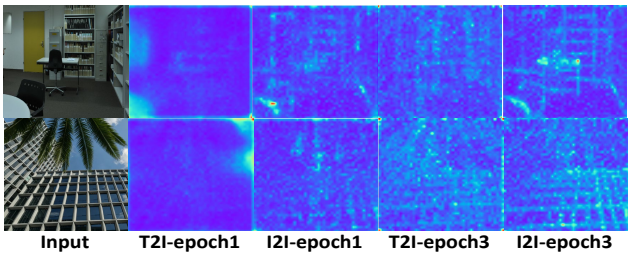


Figure 4. Visualization of Attention Map between T2I and I2I.

Effect of Pixel-Space Consistency Loss. In Section 3.2, we proposed the pixel-space consistency loss ($\mathcal{L}_{\text{Cons}}$) to bridge the gap between latent-space supervision and pixel-level quality. To verify its general effectiveness, we conducted six sets of controlled experiments across three tasks and two base models.

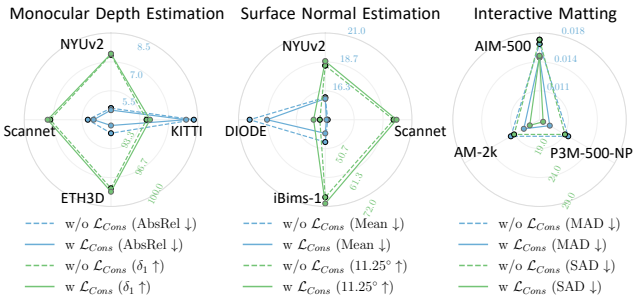


Figure 5. **Effectiveness of the Pixel-Space Consistency Loss ($\mathcal{L}_{\text{Cons}}$) across All Tasks.** The radar charts compare the performance with (solid line) and without (dashed line) our consistency loss. For each task, axes represent key metrics on different datasets (lower is better for error metrics like AbsRel, Mean, MAD, SAD; higher is better for accuracy metrics like δ_1, 11.25°).

As shown in Fig. 5 and Table 4,5, introducing $\mathcal{L}_{\text{Cons}}$ brings consistent performance improvements across all configurations. Interestingly, we observe a complementary relationship between the gains from $\mathcal{L}_{\text{Cons}}$ and the strength of the base model. For example, on the weaker T2I backbone (Table 4, IDs 1 & 2, 3 & 4), $\mathcal{L}_{\text{Cons}}$ provides substantial improvements (AbsRel drop of 1.0-1.4 on NYUv2). On the stronger I2I backbone (IDs 5 & 6, 7 & 8), where the model already possesses better structural understanding, the gains become more subtle (AbsRel drop of 0.3-0.4), indicating that $\mathcal{L}_{\text{Cons}}$ acts more as a **fine-tuner** rather than a **rectifier**. Furthermore, as illustrated in Fig. 5, this improvement is crucial for tasks highly sensitive to fine-grained structures and edges, such as normal estimation and image matting. This proves that $\mathcal{L}_{\text{Cons}}$ serves as an effective plug-and-play module that injects pixel-level geometric constraints into the latent-space generation process, thereby enhancing the final perception quality.

Effect of Theoretically Optimal Depth Mapping. In Section 3.3, we theoretically derived the optimality of the square root (Sqrt) depth mapping for minimizing relative quantization error.

We conducted four sets of controlled experiments in Table 4 (IDs 1 & 3, 2 & 4, 5 & 7, 6 & 8) to compare our Sqrt mapping against traditional uniform normalization (Uni). The results show that the Sqrt mapping yields significant performance improvements in all cases.

We observe that the improvement is particularly pronounced on the outdoor dataset KITTI, which features a larger depth range. For instance, without $\mathcal{L}_{\text{Cons}}$ (IDs 1 & 3, 5 & 7), the AbsRel reduction on KITTI (-3.0, -1.4) is much larger than on the indoor dataset NYUv2 (-0.5, -0.4). This phenomenon aligns perfectly with our theoretical analysis. As shown in formulation 10, our integral error analysis, based on the objective predicts this behavior. By substituting $g(y) = y$ (Uniform) and $g(y) = \sqrt{y}$ (Sqrt) into the integral for different depth ranges, we can quantify the theoretical error reduction. For an indoor range like

NYUv2 ($[0.1, 10]$m), the theoretical improvement for AbsRel is 0.26. For an outdoor range like KITTI ($[0.1, 80]$m), the theoretical improvement for AbsRel is a more substantial 0.6. The strong consistency between our experimental results and theoretical predictions not only proves the superiority of our depth mapping method but also highlights that a principled, first-principles-based design is crucial for unlocking the full potential of large-scale models and achieving SOTA performance.

Effect of Inference Steps. Our framework naturally supports efficient single-step inference. As shown in Fig. 6, performance peaks at 4 steps and then slightly degrades with more steps. This suggests that, unlike generative tasks, dense perception does not necessarily benefit from prolonged denoising, which may introduce over-smoothing artifacts. Despite its simplicity, single-step inference already achieves competitive performance, highlighting the efficiency of our deterministic formulation. Further details are provided in Supplementary Sections **B.2** and **C.3**.

Table 4. Ablation on the Base Model, Consistency Loss, and Depth Normalization for Monocular Depth Estimation. Here the column "D.M." stands for Depth Mapping, we compare two ways: Uni (Uniform) and Sqrt (Square Root). The best and second-best performances are highlighted.

ID	Base Model	$\mathcal{L}_{Cons}$	D.M.	NYUv2		KITTI	
				AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
1	FLUX.1		Uni	6.8	95.1	13.2	83.7
2	FLUX.1	✓	Uni	5.4	96.9	12.5	84.3
3	FLUX.1		Sqrt	6.3	95.8	10.2	89.7
4	FLUX.1	✓	Sqrt	5.3	97.0	8.4	92.8
5	FLUX.1 Kontext		Uni	5.1	96.9	9.6	91.2
6	FLUX.1 Kontext	✓	Uni	4.8	97.2	9.6	91.2
7	FLUX.1 Kontext		Sqrt	4.7	97.5	8.2	94.1
8	FLUX.1 Kontext	✓	Sqrt	4.4	97.6	7.9	94.5

Table 5. Ablation on the Base Model, Consistency Loss for Surface Normal Estimation. The best and second-best performances are highlighted.

ID	Base Model	$\mathcal{L}_{Cons}$	NYUv2		Scannet	
			Mean ↓	11.25° ↑	Mean ↓	11.25° ↑
1	FLUX.1		16.6	57.7	15.0	62.1
2	FLUX.1	✓	16.4	59.1	14.9	63.3
3	FLUX.1 Kontext		15.8	60.0	14.2	65.2
4	FLUX.1 Kontext	✓	15.7	61.6	14.1	66.3

Model Efficiency We compare our model with concurrent models (MoGe, UniDepth) in Fig. 7. Although discriminative models offer faster inference and higher precision, our approach demonstrates a significant advantage in data efficiency (requiring only about 1/100 of the data size) while achieving competitive performance, and additionally features lower FLOPs compared to other generative models. Note that we strictly adopt the **optimal ensemble configuration** (as reported in their original papers) for fair comparison: GeoWizard (10×50), Marigold v1.1 (10×4), and DepthFM (7×1).

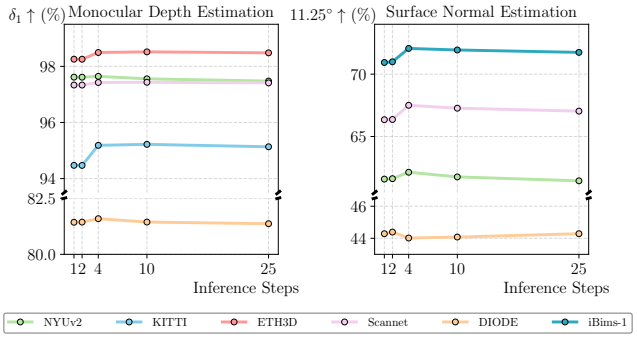


Figure 6. Ablation study on inference steps.

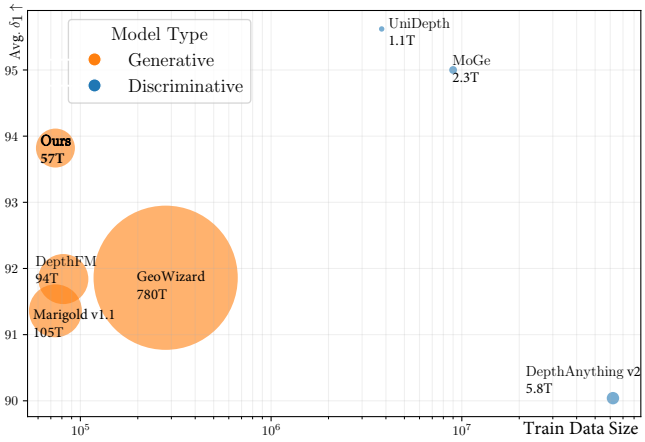


Figure 7. Comparison of Model FLOPS (Size of Bubble), Train Data Size, and Average δ_1 ↑ over 5 depth estimation benchmarks (NYUv2, KITTI, ETH3D, Scannet, DIODE).

5. Conclusion

In this work, we present **Edit2Perceive**, a simple but powerful diffusion-transformer framework for dense perception tasks. Unlike traditional text-to-image diffusion models, our approach leverages image-to-image diffusion models as geometric priors, reframing dense perception as a deterministic image-to-image transformation process. This perspective allows the model to achieve strong spatial consistency and structural fidelity across diverse tasks such as Monocular Depth Estimation, Surface Normal Estimation, and Interactive Matting. By introducing pixel-space consistency loss and efficient single-step inference, Edit2Perceive not only achieves SOTA accuracy but also significantly reduces inference cost. Our results demonstrate that diffusion-based editors can serve as a new class of perception-oriented foundation models, combining the expressive power of generative models with the precision and stability required for geometric reasoning. We believe this work opens up promising directions for unifying editing and perception within a single diffusion framework, paving the way for efficient, structure-aware, and general-purpose visual understanding.

Acknowledgment

This research is supported by the National Research Foundation, singapore under its Al singapore Programme (AlSG Award No:AISG3-RP-2022-030).

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, June 2022. 1, 2

[2] 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. 1, 2, 4

[3] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2, 3

[4] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[5] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 2

[6] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2, 3

[7] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *ICCV*, 2025. 2

[8] Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. In *NeurIPS*, 2025. 2

[9] Yan Gong, Yiren Song, Yicheng Li, Chenglin Li, and Yin Zhang. Relationadapter: Learning and transferring visual relation with diffusion transformers. *arXiv preprint arXiv:2506.02528*, 2025.

[10] Yuxin Jiang, Yuchao Gu, Yiren Song, Ivor Tsang, and Mike Zheng Shou. Personalized vision via visual in-context learning. *arXiv preprint arXiv:2509.25172*, 2025.

[11] Zitong Wang, Hang Zhao, Qianyu Zhou, Xuequan Lu, Xiangtai Li, and Yiren Song. Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. *arXiv preprint arXiv:2505.21541*, 2025.

[12] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. *arXiv preprint arXiv:2502.01572*, 2025.

[13] Yiren Song, Danze Chen, and Mike Zheng Shou. Layertracer: Cognitive-aligned layered svg synthesis via diffusion transformer. In *ICCV*, 2025.

[14] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *ICCV*, 2025.

[15] Runnan Lu, Yuxuan Zhang, Jiaming Liu, Haofan Wang, and Yiren Song. Easytext: Controllable diffusion transformer for multilingual text rendering. *arXiv preprint arXiv:2505.24417*, 2025.

[16] Wenda Shi, Yiren Song, Zihan Rao, Dengming Zhang, Jiaming Liu, and Xingxing Zou. Wordcon: Word-level typography control in scene text rendering. *arXiv preprint arXiv:2506.21276*, 2025.

[17] Wenda Shi, Yiren Song, Dengming Zhang, Jiaming Liu, and Xingxing Zou. Fonts: Text rendering with typography and style controls. In *ICCV*, 2025. 2

[18] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2022. 2

[19] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2

[20] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 2

[21] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *CVPR*, 2024. 2

[22] 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 TPAMI*, PP, 2025. 2

[23] 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. *arXiv preprint arXiv:2403.12013*, 2024. 2

[24] 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*, 2024. 2

[25] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM TOG*, 2024. 2

[26] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 2

[27] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *WACV*, 2025. 2

[28] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. *arXiv preprint arXiv:2306.05399*, 2023. 3

[29] Zixuan Ye, Wenze Liu, He Guo, Yujia Liang, Chaoyi Hong, Hao Lu, and Zhiguo Cao. Unifying automatic and interactive matting with pretrained vits. In *CVPR*, 2024. 3

[30] Longfei Huang, Yu Liang, Hao Zhang, Jinwei Chen, Wei Dong, Lunde Chen, Wanyu Liu, Bo Li, and Peng-Tao Jiang. Sdmatte: Grafting diffusion models for interactive matting. In *ICCV*, 2025. 3, 1

[31] JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, Xiangxiang Chu, and Yao Zhao. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338*, 2025. 3

[32] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, Feb 2023. 3

[33] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[34] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, 103: 102091, 2024. ISSN 1566-2535. 4

[35] Mike Roberts and Nathan Paczan. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *ICCV*, pages 10892–10902, 2020. 5

[36] Yohann Cabon, Naila Murray, and M. Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 5

[37] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*. ACM, 2022. 5

[38] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *ECCV*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012. 5

[39] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *IJCV*, 130(2):246–266, 2022. 5

[40] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *CVPR*, June 2020. 5

[41] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, pages 2970–2979, 2017. 5

[42] Ruihao Xia, Yu Liang, Peng-Tao Jiang, Hao Zhang, Qianru Sun, Yang Tang, Bo Li, and Pan Zhou. Towards natural image matting in the wild via real-scenario prior. *arXiv preprint arXiv:2410.06593*, 2024. 5

[43] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[44] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 5

[45] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *CVPR*, 2019. 5

[46] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, July 2017. 5

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

[48] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In Stefan Leal-Taixé, Laura;Roth, editor, *Eur. Conf. Comput. Vis. Worksh.*, pages 331–348. Springer International Publishing, 2019. doi: 10.1007/978-3-030-11015-4_25. 5

[49] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *ACM MM*, page 3501–3509. Association for Computing Machinery, 2021. 5

[50] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. In *IJCAI*, pages 800–806, 8 2021. 5

Edit2Perceive: Image Editing Diffusion Models Are Strong Dense Perceivers

Supplementary Material

A. Training Details

A.1. Preprocess of input RGB image

For all tasks, the input RGB image x, typically in 3-channel uint8 format, is linearly normalized to the range $[-1, 1]$ to match the VAE's input requirements.

For the Interactive Matting task, we incorporate an additional visual prompt in the form of user-provided points. Following [30], we simulate these points during training by randomly sampling up to 10 points from the foreground region. A soft mask $M_p \in [0, 1]^{H\times W}$ is then generated by placing a Gaussian kernel at each point's coordinates. This mask is also normalized to $[-1, 1]$, encoded by the VAE, and its latent representation is concatenated with the other input tokens, serving as an extra condition for the model.

A.2. Derivation of the Optimal Depth Mapping Function

This appendix provides a detailed derivation for the optimal non-linear mapping $g(y)$ that minimizes the quantization-induced relative error, as presented in Equation 10 of the main text.

Problem Formulation. The process of converting physical depth values into a model-compatible format involves four distinct stages: $y \xrightarrow{\text{mapping } g} z \xrightarrow{\text{normalization}} d \xrightarrow{\text{quantization}} q$.

Here, y represents the physical depth, $z = g(y)$ is the depth value after applying the non-linear mapping g, $d \in [-1, 1]$ is the value after normalization, and q is the final representation quantized to BF16 precision. Our objective is to determine the mapping function $g(y)$ that minimizes the relative error $\Delta y/y$ propagated back from the final quantization step.

Derivation. The BF16 (bfloat16) floating-point format uses 1 sign bit, 8 exponent bits, and 7 fraction (mantissa) bits. For a normalized value $d \in [-1, 1]$, the exponent is at most 127 (representing values up to $2^0 = 1$). The largest quantization step, Δd, for values in the range $(-1, 1)$ occurs when the exponent is 126 (for values in $[0.5, 1)$), resulting in:

$$\Delta d = 2^{(\text{exponent}-127)} \cdot 2^{-7} = 2^{(126-127)} \cdot 2^{-7} = 2^{-8} = \frac{1}{256}. \tag{14}$$

This quantization error Δd propagates backward through the preceding stages. The linear normalization maps the range of the function g, denoted as $[z_{\min}, z_{\max}] = [g(y_{\min}), g(y_{\max})]$, to the interval $[-1, 1]$. The normalization is defined as $d = \frac{z-z_{\min}}{z_{\max}-z_{\min}} \cdot 2 - 1$. The error in the mapped space, Δz, is therefore:

$$\Delta z = \frac{z_{\max} - z_{\min}}{2}\Delta d = \frac{g(y_{\max}) - g(y_{\min})}{512}. \tag{15}$$

Using the chain rule, we can express the error in the original physical depth space, Δy, as $\Delta y \approx \frac{dy}{dz}\Delta z$. Since $z = g(y)$, we have $\frac{dz}{dy} = g'(y)$, which implies $\frac{dy}{dz} = \frac{1}{g'(y)}$. Our goal is to minimize the relative error, $\Delta y/y$, across the entire depth range $[y_{\min}, y_{\max}]$. The error at any given point y is:

$$\frac{\Delta y}{y} = \frac{1}{y}\frac{dy}{dz}\Delta z = \frac{1}{y \cdot g'(y)}\frac{g(y_{\max}) - g(y_{\min})}{512}. \tag{16}$$

To find the optimal function g that minimizes this error over the continuous range, we formulate the problem as the minimization of the average relative error, which is equivalent to minimizing its integral:

$$\min_g \int_{y_{\min}}^{y_{\max}} \frac{1}{y \cdot g'(y)}\mathrm{d}y \cdot [g(y_{\max}) - g(y_{\min})]. \tag{17}$$

(Note: This expression, up to a constant factor, is what is presented in Equation 10).

We can rewrite the term $g(y_{\max}) - g(y_{\min})$ as the integral of its derivative: $\int_{y_{\min}}^{y_{\max}} g'(y)\mathrm{d}y$. Substituting this into our objective function gives:

$$\min_g \left(\int_{y_{\min}}^{y_{\max}} g'(y)\mathrm{d}y\right)\left(\int_{y_{\min}}^{y_{\max}} \frac{1}{y \cdot g'(y)}\mathrm{d}y\right). \tag{18}$$

This expression is in the form of the product of two integrals, which can be addressed using the Cauchy-Schwarz inequality for integrals. The inequality states that for any two functions $A(y)$ and $B(y)$: $(\int A(y)B(y)\mathrm{d}y)^2 \leq (\int A(y)^2\mathrm{d}y)(\int B(y)^2\mathrm{d}y)$.

Let's define $A(y) = \sqrt{g'(y)}$ and $B(y) = \frac{1}{\sqrt{y \cdot g'(y)}}$. Then:

- $\int A(y)^2\mathrm{d}y = \int g'(y)\mathrm{d}y$
- $\int B(y)^2\mathrm{d}y = \int \frac{1}{y \cdot g'(y)}\mathrm{d}y$

The product of these two integrals is minimized when the equality in the Cauchy-Schwarz inequality holds. This occurs if and only if one function is a constant multiple of the other, i.e., $A(y) = k \cdot B(y)$ for some constant k.

$$\sqrt{g'(y)} = k \cdot \frac{1}{\sqrt{y \cdot g'(y)}} \tag{19}$$

$$\implies g'(y) = \frac{k^2}{y \cdot g'(y)} \tag{20}$$

$$\implies (g'(y))^2 = \frac{k^2}{y} \tag{21}$$

$$\implies g'(y) \propto \frac{1}{\sqrt{y}}. \tag{22}$$

Integrating this result with respect to y yields the optimal form for the mapping function $g(y)$:

$$g(y) \propto \sqrt{y}. \tag{23}$$

This derivation proves that a square-root mapping is theoretically optimal for minimizing the relative quantization error when representing depth values under BF16 precision.

A.3. Derivation of the Numerically Stable Normal Consistency Loss

Our pixel-space consistency loss for surface normal estimation, as presented in Section 3.2, is designed for numerical stability during training. A naive approach to compute the mean angular error between the ground-truth normal y and the predicted normal $\hat{y}$ (assuming both are unit vectors) is to use the arccosine function:

$$\mathcal{L}_{\text{naive}} = \mathbb{E}\left[\arccos(y \cdot \hat{y})\right]. \tag{24}$$

However, the derivative of the *arccos* function is given by:

$$\frac{d}{dx}\arccos(x) = -\frac{1}{\sqrt{1-x^2}}. \tag{25}$$

As the argument $x = y \cdot \hat{y}$ approaches ± 1 (i.e., when the predicted normal is very accurate and nearly collinear with the ground truth), the denominator of Eq. 25 approaches zero. This causes the gradient to explode, leading to numerical instability and training divergence.

To circumvent this issue, we adopt a more robust formulation based on the two-argument arctangent function, *atan2*. The angle θ between two unit vectors can be uniquely determined by its sine and cosine values, which correspond to the magnitude of their cross product and their dot product, respectively:

$$\sin(\theta) = \|y \times \hat{y}\|_2, \tag{26}$$

$$\cos(\theta) = y \cdot \hat{y}. \tag{27}$$

Our final loss, as used in the main paper, is then formulated as:

$$\mathcal{L}_{\text{Cons}}^{\text{normal}} = \mathbb{E}\left[\text{atan2}\left(\|y \times \hat{y}\|_2, \quad y \cdot \hat{y}\right)\right]. \tag{28}$$

The *atan2* function has well-defined, bounded gradients across its entire domain, which resolves the instability issue and significantly improves training stability for the normal estimation task.

B. Additional Quantitative Results

This appendix provides the complete quantitative results for the ablation studies discussed in the main paper.

B.1. Detailed Ablation Studies

For brevity, the main paper analyzes the impact of the base model, consistency loss, and depth mapping on a subset of datasets. Here, we present the full results across all benchmark datasets.

Tables 6, 7, and 8 provide the comprehensive ablation results for monocular depth estimation, surface normal estimation, and interactive matting, respectively. These tables serve as a supplement to Tables 4, 5, and Figure 5 in the main text.

The complete results confirm the conclusions drawn in the main paper: (i) the I2I-based model (FLUX.1 Kontext) consistently outperforms the T2I-based model (FLUX.1); (ii) the pixel-space consistency loss ($\mathcal{L}_{\text{Cons}}$) brings universal performance improvements; (iii) our theoretically optimal square root depth mapping (Sqrt) is significantly superior to uniform normalization (Uni).

B.2. Analysis of Inference Steps

The complete results for the analysis of inference steps are provided in Tables 9, 10, and 11. Our framework demonstrates highly efficient inference capabilities. For all experiments reported in the main paper, we use single-step inference by default. As shown in the tables, the performance degradation from using a single step compared to multiple steps is minor and acceptable, confirming the effectiveness of our efficient approach.

C. Additional Qualitative Results

C.1. Comparison of other SOTA models

Figure 8 provides additional qualitative comparisons for zero-shot monocular depth estimation. We observe that our model, Edit2Perceive, demonstrates superior performance in capturing complex scene geometry compared to prior works. For instance, our method accurately reconstructs fine-grained details such as the folds of the curtains (second and fourth columns) and the intricate structure of pine needles within shadowed regions (first column), highlighting its powerful capability for detailed geometric reasoning.

In Figure 9, we present further qualitative comparisons for zero-shot surface normal estimation. Our model excels in scenarios with complex and subtle textures. Notably, it successfully captures the rough texture of the tree bark and the delicate structure of leaves (second column), as well as the fine surface patterns on the backpack (third column). This demonstrates the model's robustness in recovering detailed surface geometry from challenging in-the-wild images.

Figure 10 illustrates the superior performance of our model on the interactive matting task. Edit2Perceive exhibits exceptional capability in handling extremely fine details and challenging materials. It accurately delineates delicate structures like feathers and hair, and correctly handles semi-transparent objects such as glass cups and water droplets, setting it apart from competing methods.

C.2. Visual Ablation Study of Components

To visually dissect the contribution of each component, we present the ablation results for depth estimation in Fig-

Table 6. Additional Ablation Study on the Base Model, Consistency Loss, and Depth Normalization for Monocular Depth Estimation. Here the column "D.M." stands for Depth Mapping, we compare two ways: Uni (Uniform) and Sqrt (Square Root). The best and second-best performances are highlighted.

ID	Base Model	$\mathcal{L}_{Cons}$	D.M.	NYUv2		KITTI		ETH3D		Scannet		DIODE	
				AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
1	Flux.1		Uni	6.8	95.1	83.7	13.2	7.4	94.7	8.3	92.7	30.1	77.2
2	Flux.1	✓	Uni	5.4	96.9	84.3	12.5	6.3	95.4	6.1	96.2	29.3	77.5
3	Flux.1		Sqrt	6.3	95.8	89.7	10.2	6.3	96.6	7.5	93.8	26.4	78.9
4	Flux.1	✓	Sqrt	5.3	97.0	92.8	8.4	5.7	97.1	6.5	95.9	25.5	79.8
5	Flux.1 Kontext		Uni	5.1	96.9	91.2	9.6	5.4	96.5	5.2	96.8	29.2	78.9
6	Flux.1 Kontext	✓	Uni	4.8	97.2	91.2	9.6	5.3	96.9	5.3	96.7	28.9	79.3
7	Flux.1 Kontext		Sqrt	4.7	97.5	94.1	8.2	4.7	98.0	5.3	97	25.2	81.0
8	Flux.1 Kontext	✓	Sqrt	4.4	97.6	94.5	7.9	4.3	98.3	4.9	97.3	24.8	81.4

Table 7. Additional Ablation Study on the Base Model, Consistency Loss for Surface Normal Estimation. The best and second-best performances are highlighted.

ID	Base Model	$\mathcal{L}_{Cons}$	NYUv2		Scannet		iBims-1		DIODE	
			Mean ↓	11.25° ↑	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑
1	Flux.1		16.6	57.7	15.0	62.1	17.0	65.1	19.9	44.7
2	Flux.1	✓	16.4	59.1	14.9	63.3	16.8	66.2	19.9	40.1
3	Flux.1 Kontext		15.8	60.0	14.2	65.2	15.8	68.6	20.1	42.0
4	Flux.1 Kontext	✓	15.7	61.6	14.1	66.3	15.1	70.9	18.7	44.3

Table 8. Additional Ablation Study on the Base Model, Consistency Loss for Interactive Matting. The best and second-best performances are highlighted.

ID	Base Model	$\mathcal{L}_{Cons}$	AIM-500					P3M-500-NP					AM-2k				
			MSE↓	MAD↓	SAD↓	Grad↓	Conn↓	MSE↓	MAD↓	SAD↓	Grad↓	Conn↓	MSE↓	MAD↓	SAD↓	Grad↓	Conn↓
1	FLUX.1		0.0495	0.085	144.25	23.74	72.02	0.0316	0.069	108.94	35.80	61.29	0.011	0.027	46.26	14.08	27.59
2	FLUX.1	✓	0.0490	0.084	142.23	23.54	70.31	0.0299	0.066	104.22	35.91	59.78	0.0102	0.024	45.13	13.58	25.87
3	FLUX.1 Kontext		0.0058	0.017	30.84	18.12	17.51	0.0034	0.011	22.64	10.81	12.85	0.0039	0.012	21.53	9.81	12.85
4	FLUX.1 Kontext	✓	0.0057	0.017	29.14	18.28	15.73	0.0028	0.011	19.39	13.21	10.17	0.0037	0.012	20.42	9.61	9.94

Table 9. Additional Ablation Study on the Inference Steps of Monocular Depth Estimation task. The best and second-best performances are highlighted.

Inference Steps	NYUv2		KITTI		ETH3D		Scannet		DIODE	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
1	4.425	97.616	7.913	94.472	4.283	98.256	4.886	97.334	24.848	81.429
2	4.423	97.614	7.910	94.470	4.281	98.254	4.888	97.333	24.846	81.436
4	4.198	97.644	7.375	95.185	3.637	98.494	4.602	97.425	24.846	81.589
10	4.264	97.555	7.417	95.220	3.631	98.517	4.592	97.433	25.026	81.438
25	4.312	97.475	7.487	95.134	3.669	98.482	4.615	97.411	25.145	81.365

Table 10. Ablation on the Inference Steps for Surface Normal Estimation. The best and second-best performances are highlighted.

Inference Steps	NYUv2		Scannet		iBims-1		DIODE	
	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑	Mean ↓	11.25° ↑
1	15.668	61.584	14.105	66.347	15.136	70.936	18.710	44.287
2	15.663	61.619	14.101	66.365	15.119	70.991	18.683	44.389
4	16.239	62.134	14.254	67.503	14.923	72.063	18.507	44.019
10	16.627	61.761	14.599	67.270	15.183	71.938	18.628	44.079
25	16.830	61.445	14.752	67.035	15.344	71.746	18.710	44.287

ure 11. **Base Model:** Comparing models with identical settings but different base models (e.g., ID 1 vs. 5, ID 2 vs. 6, etc.), it is evident that the FLUX.1 Kontext (I2I) based models (IDs 5-8) consistently produce more accurate and structurally sound results than their FLUX.1 (T2I) counterparts (IDs 1-4). **Consistency Loss:** The effect of our pixel-space consistency loss can be seen by comparing adjacent columns (e.g., ID 1 vs. 2, ID 5 vs. 6). The addition of $\mathcal{L}_{\text{Cons}}$ (IDs 2,4,6,8) consistently enhances fine-grained details, as highlighted by the sharper reconstruction of the cur-

Table 11. Additional Ablation Study on the Base Model, Consistency Loss for Interactive Matting. The best and second-best performances are highlighted.

Inference Steps	AIM-500					P3M-500-NP					AM-2k				
	MSE↓	MAD↓	SAD↓	Grad↓	Conn↓	MSE↓	MAD↓	SAD↓	Grad↓	Conn↓	MSE↓	MAD↓	SAD↓	Grad↓	Conn↓
1	0.0057	0.017	29.14	18.28	15.73	0.0028	0.011	19.39	13.21	10.17	0.0037	0.012	20.42	9.61	9.94
2	0.0056	0.017	28.32	18.16	15.35	0.0028	0.011	19.13	13.01	10.06	0.0034	0.012	20.16	9.59	9.93
4	0.0055	0.013	21.97	16.14	13.66	0.0022	0.006	11.14	12.17	8.04	0.0032	0.008	12.74	8.52	7.99
10	0.0059	0.013	21.74	16.59	13.34	0.0025	0.006	10.92	12.54	7.63	0.0034	0.008	12.66	8.95	7.96
25	0.0059	0.014	23.48	17.32	13.20	0.0029	0.008	13.37	13.25	7.65	0.0034	0.009	14.74	9.59	8.10

tains (indicated by arrows). **Depth Mapping:** Comparing different depth normalization methods (e.g., ID 1 vs. 3, ID 2 vs. 4), our proposed Sqrt mapping (IDs 3,4,7,8) yields visibly superior results compared to the Uniform mapping (IDs 1,2,5,6), particularly in preserving depth variations.

Figure 12 visualizes the ablation study for surface normal estimation. We observe that without the consistency loss (IDs 1 & 3), the predictions are prone to speckled artifacts and noisy patterns. The introduction of our pixel-space supervision, $\mathcal{L}_{\text{Cons}}$, (IDs 2 & 4) significantly mitigates these issues, resulting in much smoother and more coherent normal maps. Furthermore, comparing the base models, FLUX.1 Kontext (IDs 3 & 4) demonstrates a markedly improved ability to discern complex edges compared to FLUX.1 (IDs 1 & 2).

C.3. Comparison of Inference Steps

Figures 13 and 14 visualize the effect of varying the number of inference steps for depth and normal estimation, respectively. We observe that while additional steps can offer marginal refinements in edge sharpness, our single-step inference already produces high-quality and structurally coherent results. The performance gain from multi-step inference is minimal, confirming that our approach offers an excellent trade-off between efficiency and quality with negligible and acceptable performance loss.

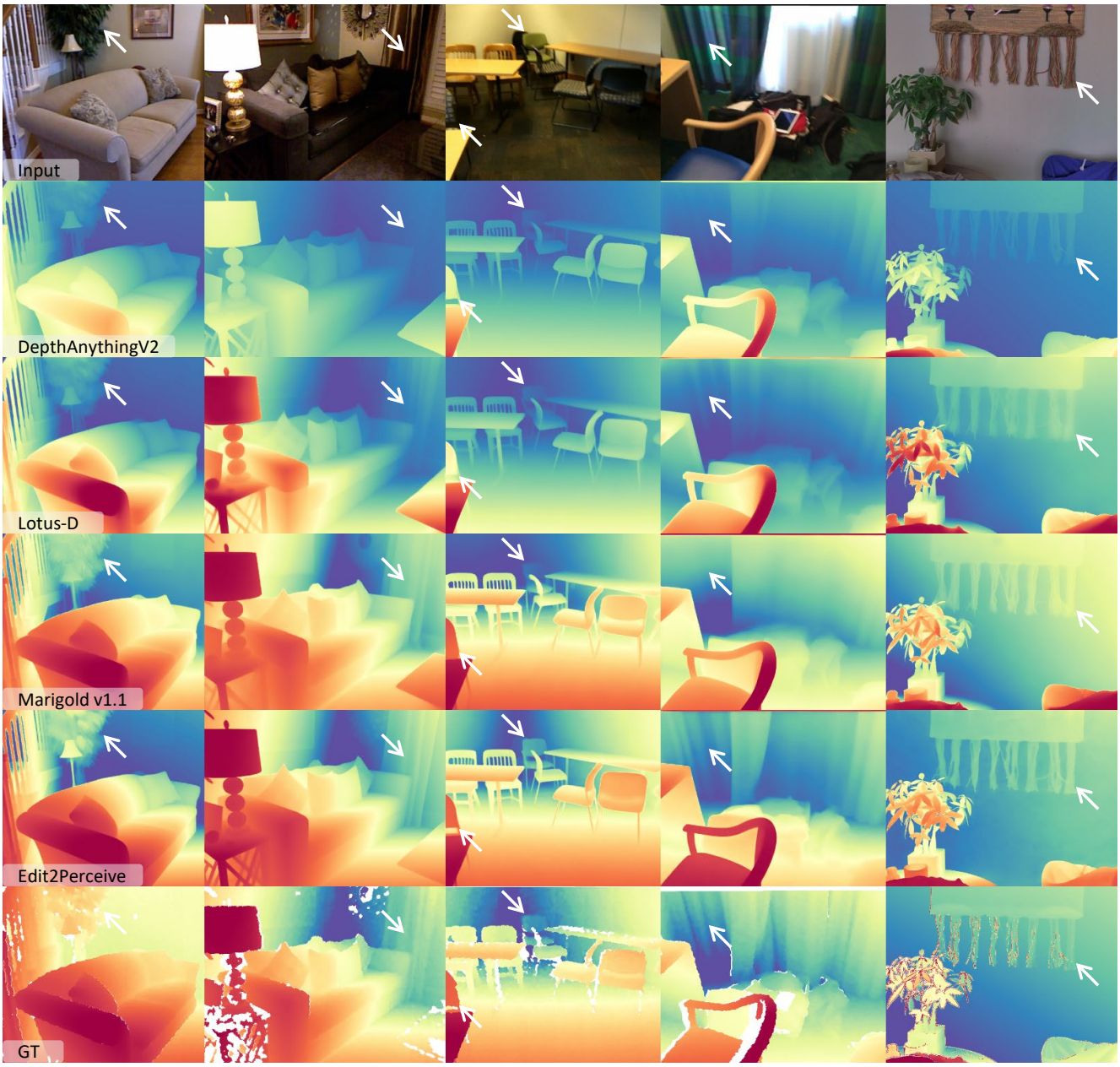


Figure 8. **Additional Qualitative Comparisons for Zero-Shot Monocular Depth Estimation.** Our method consistently produces more detailed and structurally coherent depth maps compared to other state-of-the-art methods across a variety of challenging indoor and outdoor scenes.

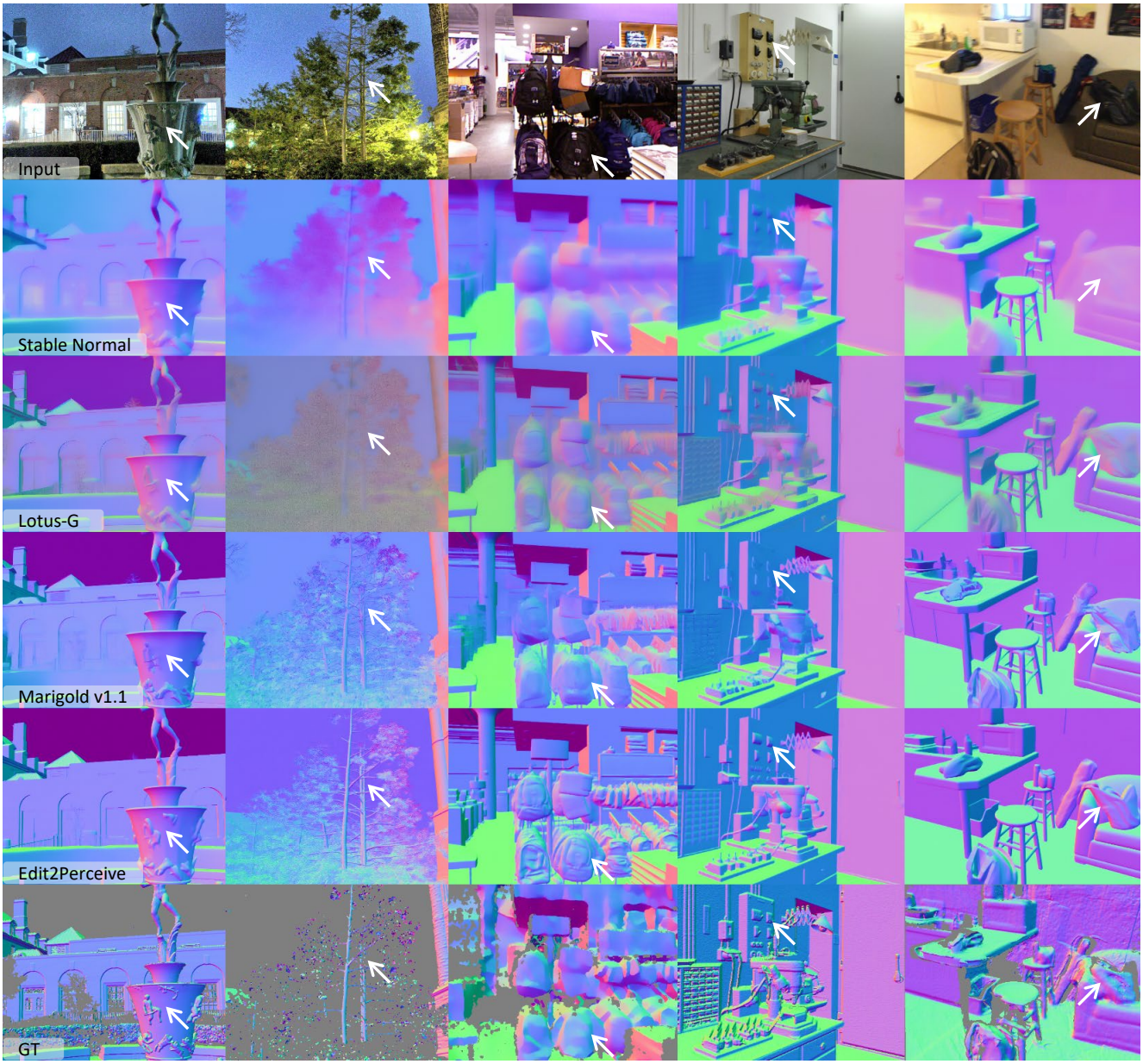


Figure 9. **Additional Qualitative Comparisons for Zero-Shot Surface Normal Estimation.** Compared to other methods, our model demonstrates a superior ability to capture fine-grained surface details and subtle curvatures, such as the texture of tree bark (second column) and fabric patterns (third column).

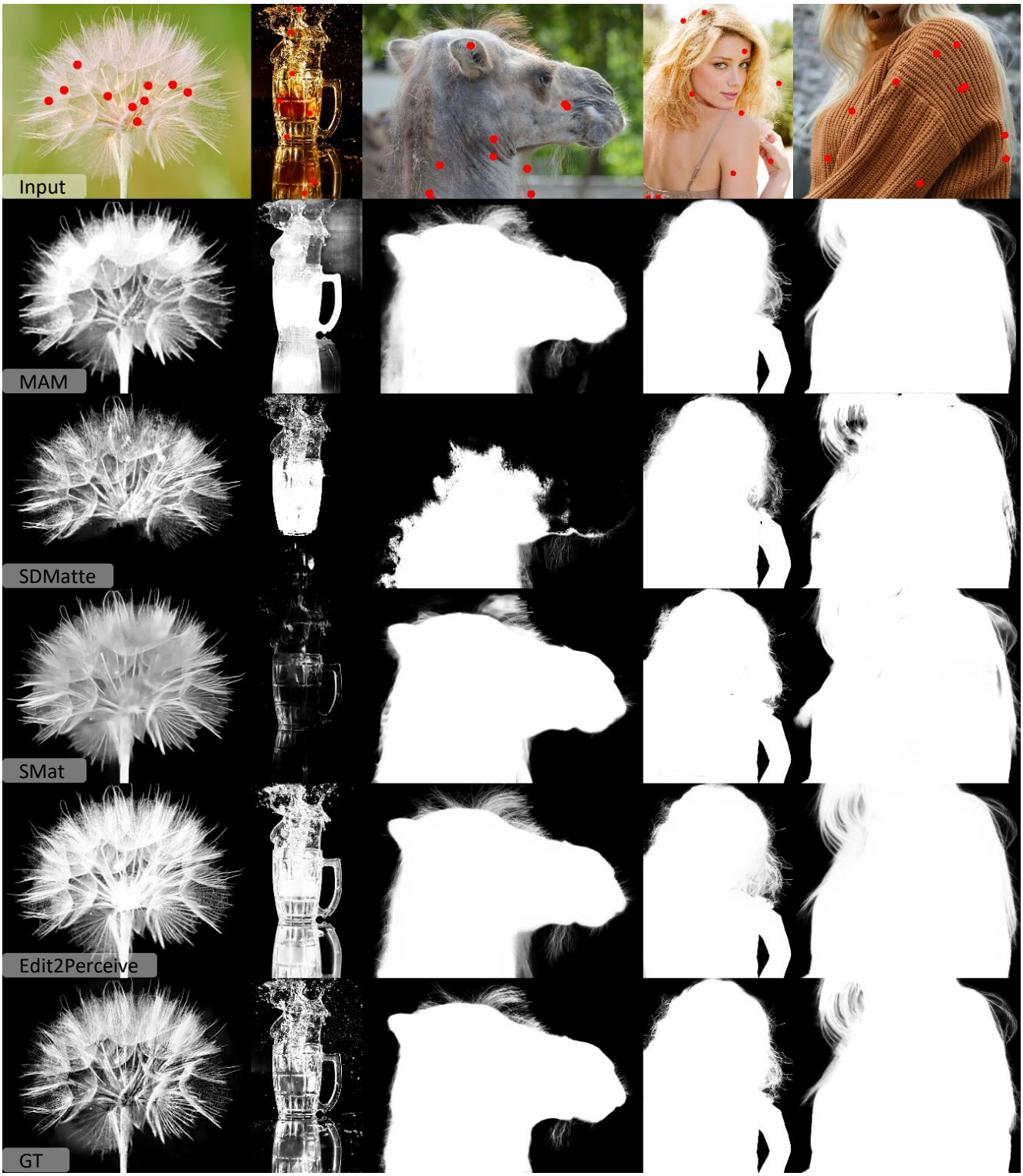


Figure 10. **Additional Qualitative Comparisons for Interactive Matting.** Our method excels at handling challenging cases, including extremely fine structures like hair and feathers, as well as semi-transparent materials like glass and water droplets, producing significantly cleaner and more accurate alpha mattes.

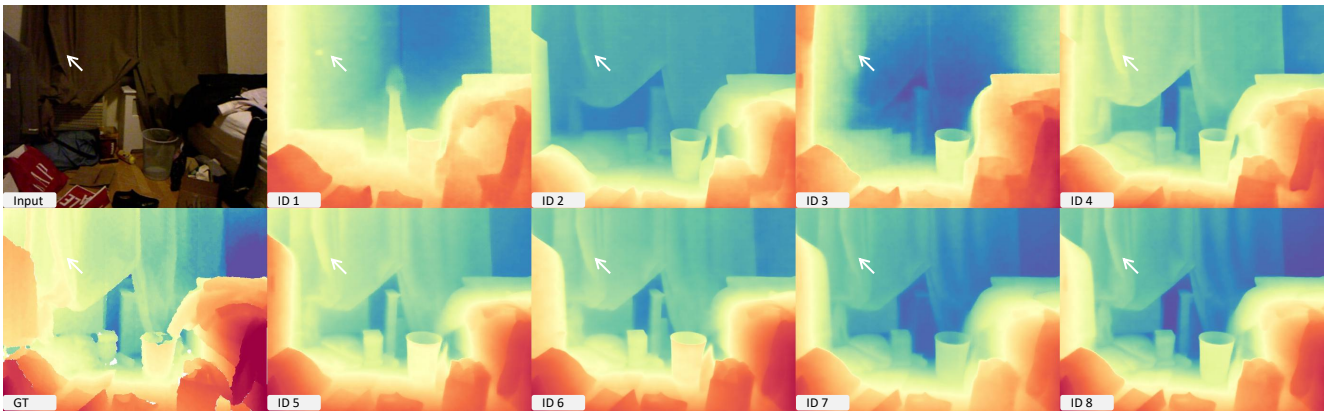


Figure 11. **Visual Ablation Study for Monocular Depth Estimation.** Each ID corresponds to ID in Table 6, allowing direct visual assessment of each component's impact. These results visually confirm the quantitative findings: the I2I base model, the consistency loss, and our Sqrt depth mapping each contribute significantly to the final performance.

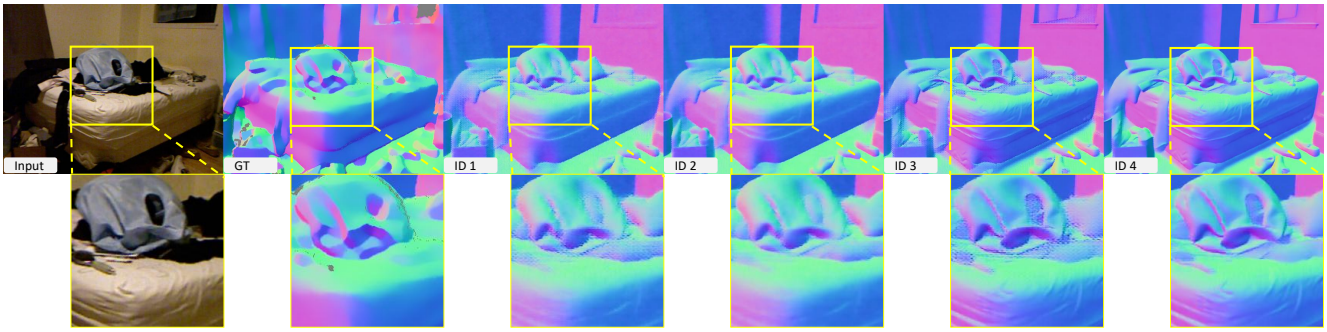


Figure 12. **Visual Ablation Study for Surface Normal Estimation.** Each column corresponds to an ID from Table 7. The zoomed-in regions (below) highlight how our consistency loss effectively removes speckled artifacts (ID 1 vs. 2 and 3 vs. 4) and how the I2I base model better captures complex geometry (ID 1-2 vs. 3-4).

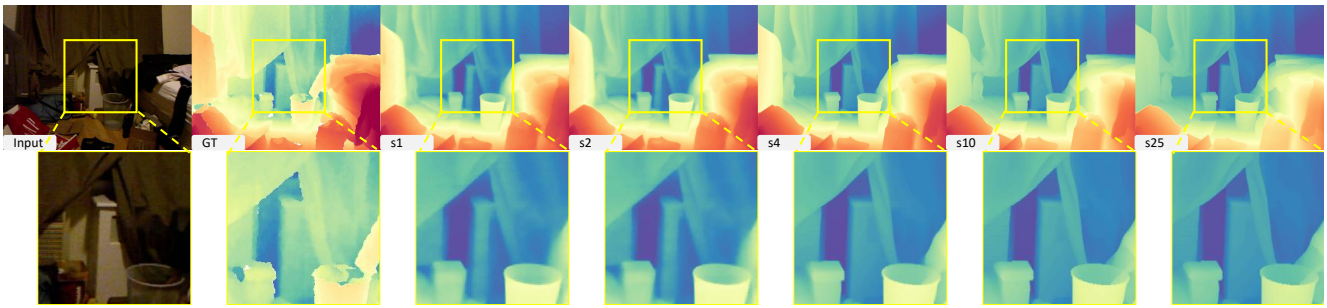


Figure 13. **Visualizing the Effect of Inference Steps on Depth Estimation.** The zoomed-in regions (below) show that while increasing the number of steps from 1 to 4 offers slight improvements in detail, further steps yield diminishing returns. This demonstrates that our single-step inference already achieves high-quality results.

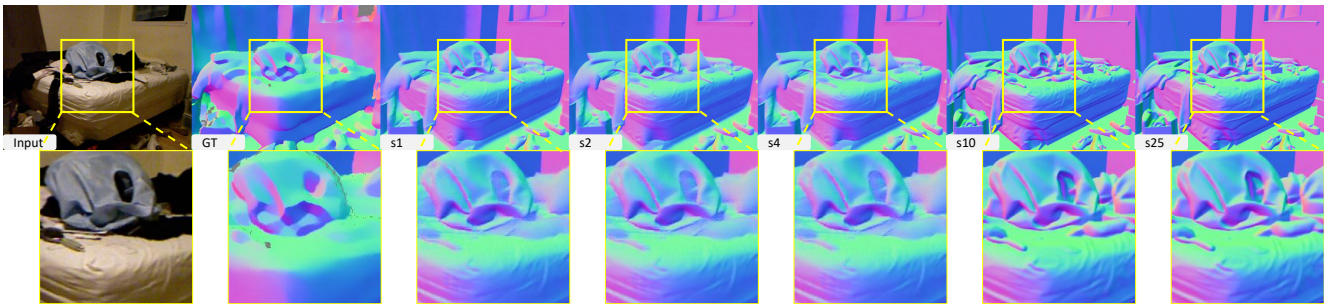


Figure 14. **Visualizing the Effect of Inference Steps on Normal Estimation.** Similar to depth estimation, we observe that single-step inference produces results comparable to multi-step inference. The performance gain from additional steps is marginal, highlighting the efficiency of our method.