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.

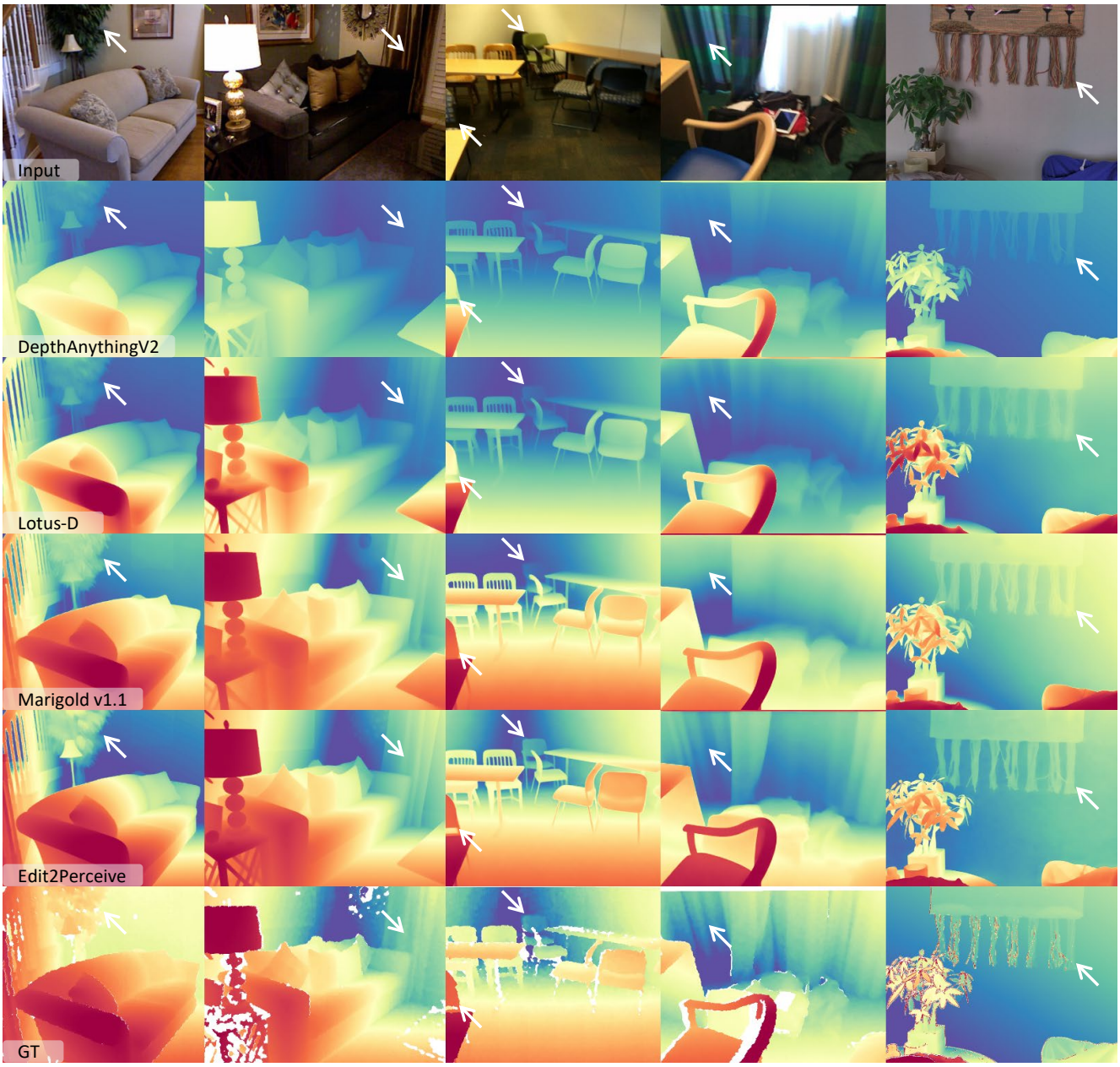

Figure 8. **Additional Qualitative Comparisons for Zero-Shot Monocular Depth Estimation.** Our method consistently produces more detailed and structurally coherent depth maps compared to other state-of-the-art methods across a variety of challenging indoor and outdoor scenes.

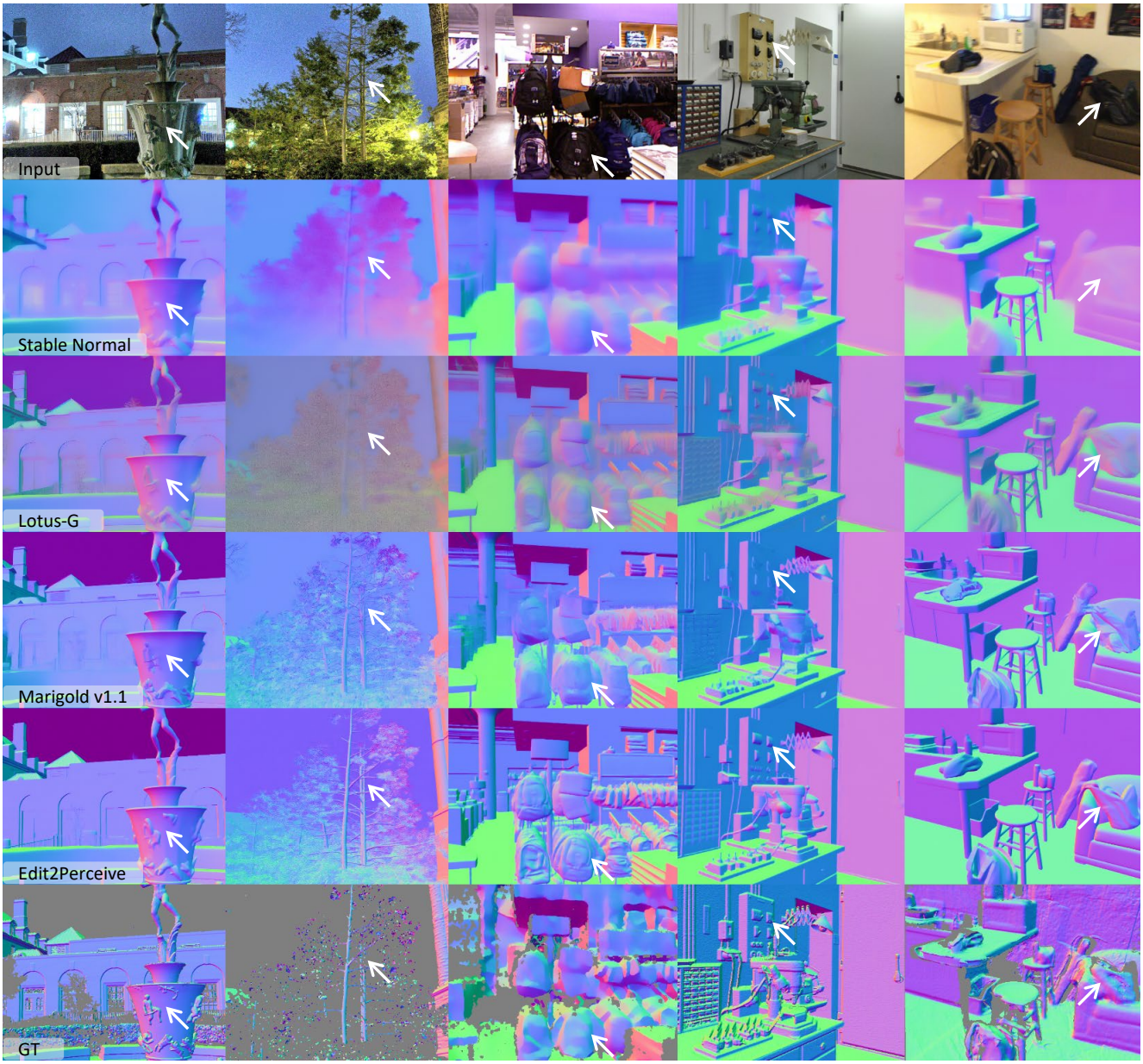

Figure 9. **Additional Qualitative Comparisons for Zero-Shot Surface Normal Estimation.** Compared to other methods, our model demonstrates a superior ability to capture fine-grained surface details and subtle curvatures, such as the texture of tree bark (second column) and fabric patterns (third column).

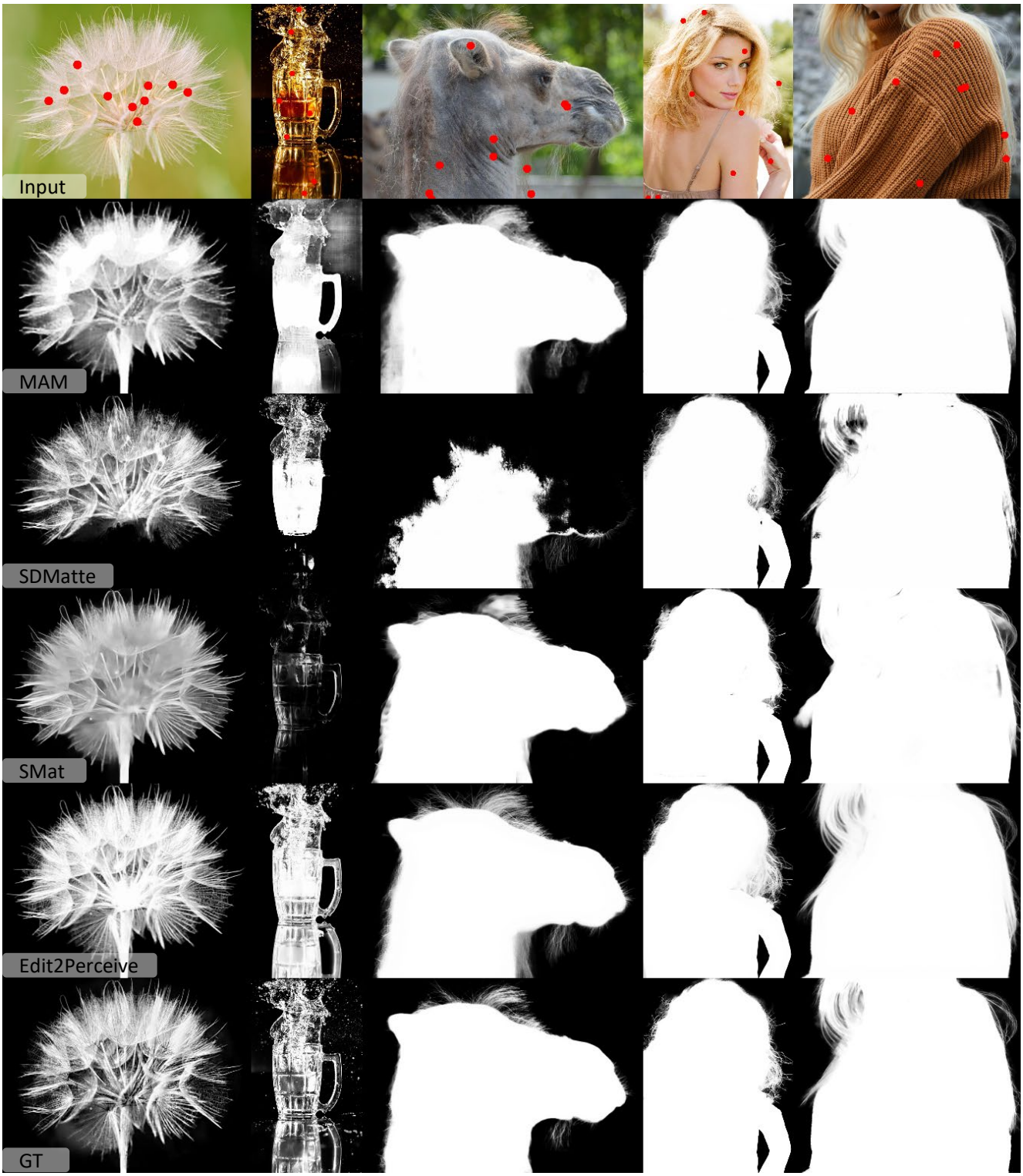

Figure 10. **Additional Qualitative Comparisons for Interactive Matting.** Our method excels at handling challenging cases, including extremely fine structures like hair and feathers, as well as semi-transparent materials like glass and water droplets, producing significantly cleaner and more accurate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

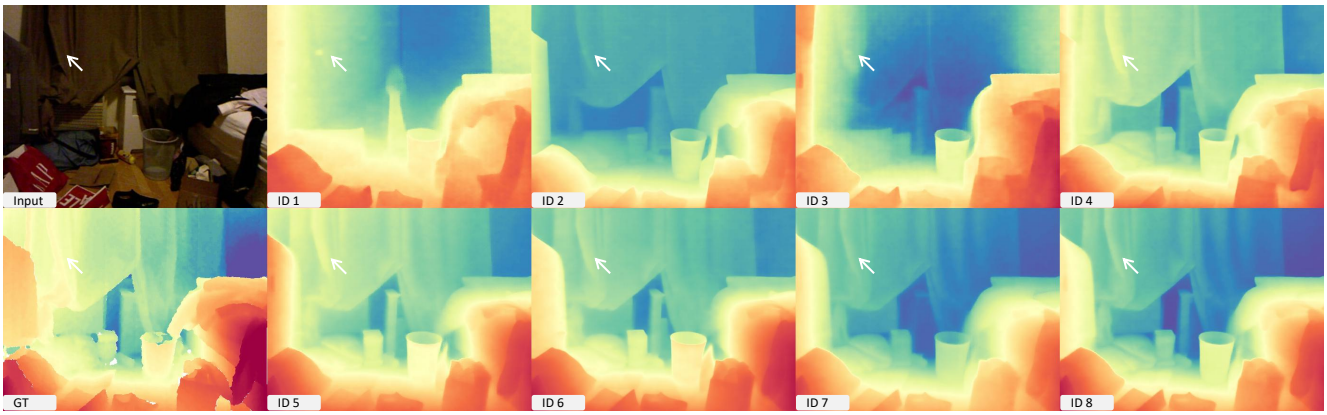

Figure 11. **Visual Ablation Study for Monocular Depth Estimation.** Each ID corresponds to ID in Table 6, allowing direct visual assessment of each component's impact. These results visually confirm the quantitative findings: the I2I base model, the consistency loss, and our Sqrt depth mapping each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

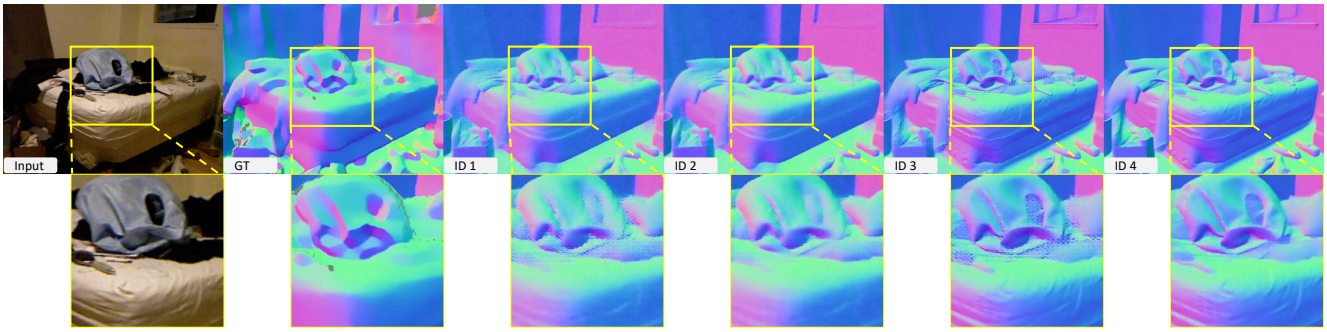

Figure 12. **Visual Ablation Study for Surface Normal Estimation.** Each column corresponds to an ID from Table 7. The zoomed-in regions (below) highlight how our consistency loss effectively removes speckled artifacts (ID 1 vs. 2 and 3 vs. 4) and how the I2I base model better captures complex geometry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

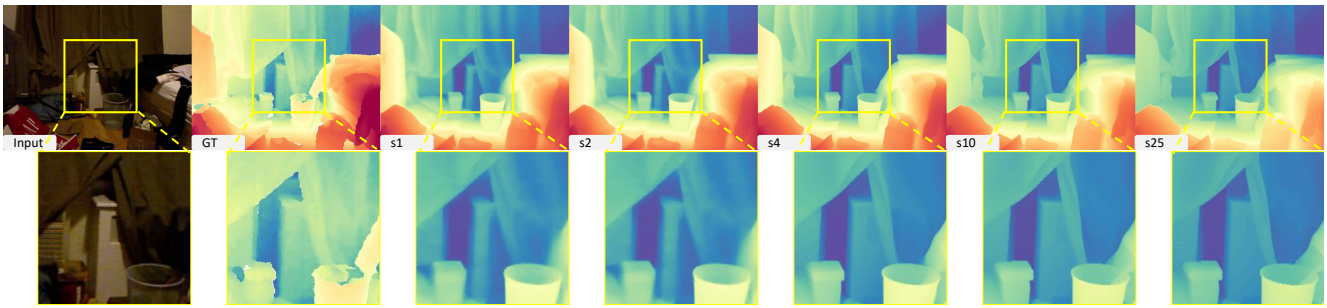

Figure 13. **Visualizing the Effect of Inference Steps on Depth Estimation.** The zoomed-in regions (below) show that while increasing the number of steps from 1 to 4 offers slight improvements in detail, further steps yield diminishing returns. This demonstrates that our single-step inference already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

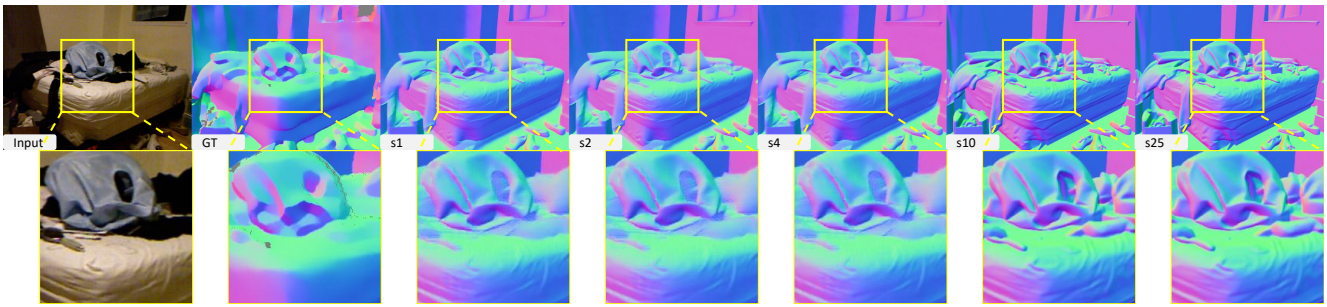

Figure 14. **Visualizing the Effect of Inference Steps on Normal Estimation.** Similar to depth estimation, we observe that single-step inference produces results comparable to multi-step inference. The performance gain from additional steps is marginal, highlighting the efficiency of our method.