Scaling Properties of Diffusion Models For Perceptual Tasks

Supplementary Material

1. Scaling Power Law Derivation

We derive all scaling laws in our figures using an iterative method based on the convex hull. This procedure ensures the curve accurately captures the minimal envelope of the data, representing the scaling behavior of loss as a function of computational cost. The algorithm begins by aggregating all data points representing the relationship between the loss and MACs. We compute the convex hull of the aggregated points, which forms the smallest convex boundary enclosing all points. From this hull, the lower envelope is extracted. These lower hull points represent the minimal set of points along the loss vs. MACs curve, which define the primary trend. The scaling law is modeled as:

$$L(C) = a \times C^b, \tag{1}$$

where $L(C)$ is the loss/error, C represents the compute in MACs, and a, b are parameters to be optimized. The fitting process is initialized with reasonable guesses for these parameters and constraints to ensure the solution remains physically meaningful (e.g., non-negative losses). After fitting the initial curve to the lower hull points, the method identifies any data points that lie below the fitted curve. These points indicate regions where the current fit does not fully encapsulate the minimal envelope of the data. These points are added to the lower hull, and the convex hull is recalculated to include them. The fitting process is repeated iteratively until convergence, where either fewer than N_{max} points are found below the fitted curve or a maximum number of iterations is reached. This iterative process ensures the scaling law curve fully captures the trend defined by the lower envelope of the data. The final parameters a and b are determined after convergence, and the resulting curve represents the optimal scaling power law for the loss/error vs. compute relationship.

2. Noise Variance Schedule Visualization

During the denoising process, our mixture-of-experts generalist model refines depth latents from timestep $t = 1000$ to $t = 0$. At selected timesteps ($t \in \{1000, 800, 600, 400, 200, 0\}$), we project the current denoised depth latent into RGB space and compress the representation along the channel dimension to retrieve depth predictions.

To align these predictions with the ground truth, we apply least squares regression at each timestep to determine scaling and shifting parameters, γ and β, respectively. These parameters are used to scale and shift the predictions, ensuring consistency with the ground truth depth. Fig. 1 illustrates the progression of the denoising process, with the predictions approaching the ground truth as $t \to 0$.

3. Additional Results From Generalist Model

We visualize additional samples from our mixture-of-experts generalist model for depth estimation, optical flow estimation, and amodal segmentation in Fig. 2. Our model is able to generalize across the three tasks with accurate visual results, displaying the effectiveness of our scaling techniques to train a generalist diffusion model for perception.

4. Evaluation Metrics

We use a variety of metrics to evaluate our models. For depth estimation, we use Delta1 Accuracy and Absolute Relative Error metrics. The δ_1 accuracy measures the percentage of predicted depth values where the ratio between the prediction and ground truth (or its inverse) is within a threshold. The absolute relative error quantifies the mean of the absolute difference between the predicted and ground truth depths relative to the ground truth.

$$\delta_1 = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\max\left(\frac{\hat{D}_i}{D_i}, \frac{D_i}{\hat{D}_i}\right) < 1.25\right), \tag{2}$$

$$\text{AbsRel} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|\hat{D}_i - D_i\right|}{D_i}. \tag{3}$$

For optical flow estimation, we measure end-point error, which measures the average Euclidean distance between the predicted flow vectors and the ground truth flow vectors.

$$\text{EPE} = \frac{1}{N}\sum_{i=1}^{N} \left\|\hat{\mathbf{F}}_i - \mathbf{F}_i\right\|_2, \tag{4}$$

Finally, for amodal segmentation, we evaluate our model by computing the mIOU, calculated as the average IoU over all samples. IoU provides an intuitive measure of the overlap between the predicted segmentation and the ground truth, with values ranging from 0 (no overlap) to 1 (perfect overlap).

$$\text{IoU}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{5}$$

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU}_i \tag{6}$$

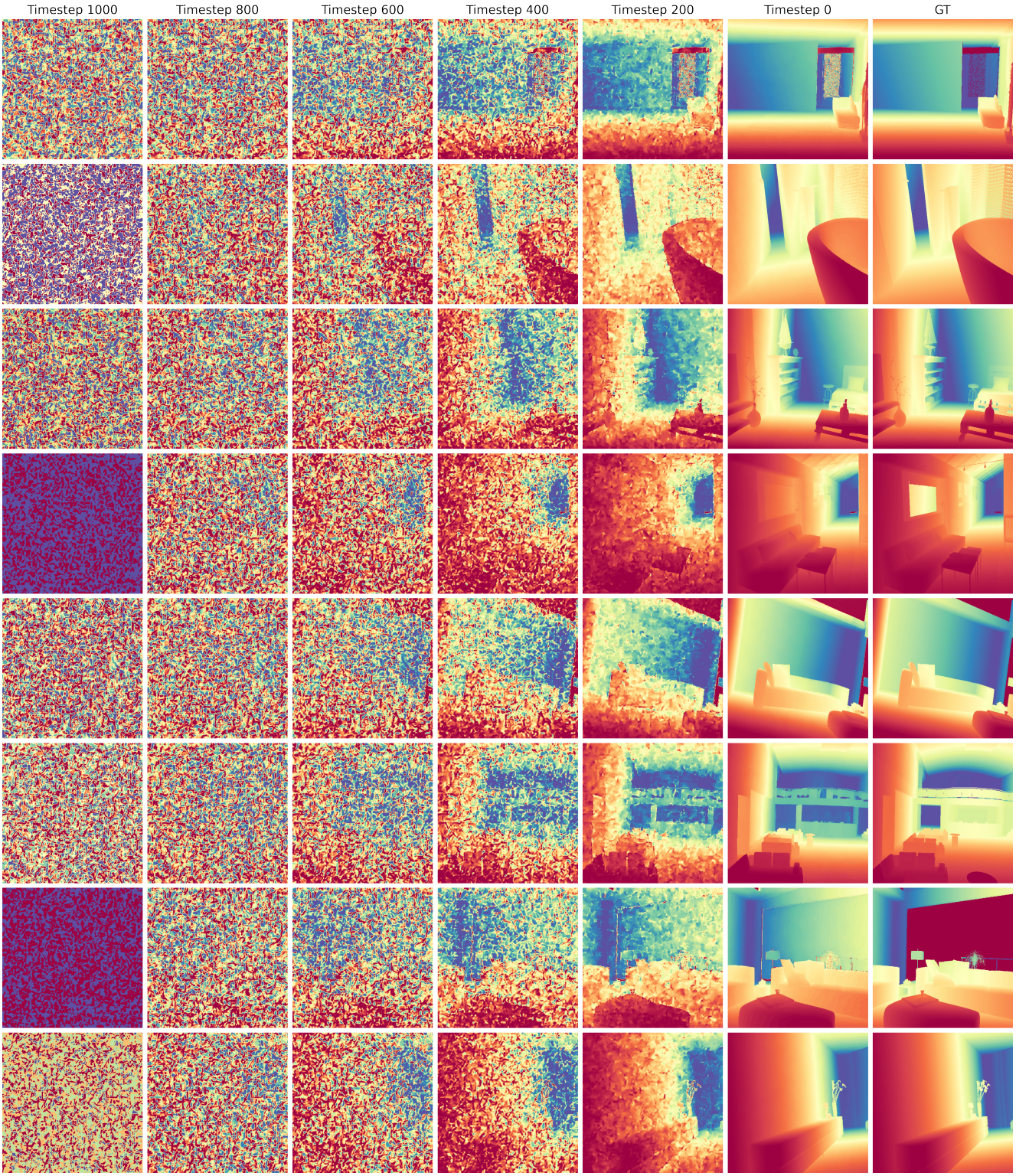

Figure 1. **Noise Variance Schedule Progression:** We project depth latents at uniform timesteps in the denoising process to show the predicted depth maps. The samples in this figure are generated from the Hypersim dataset.

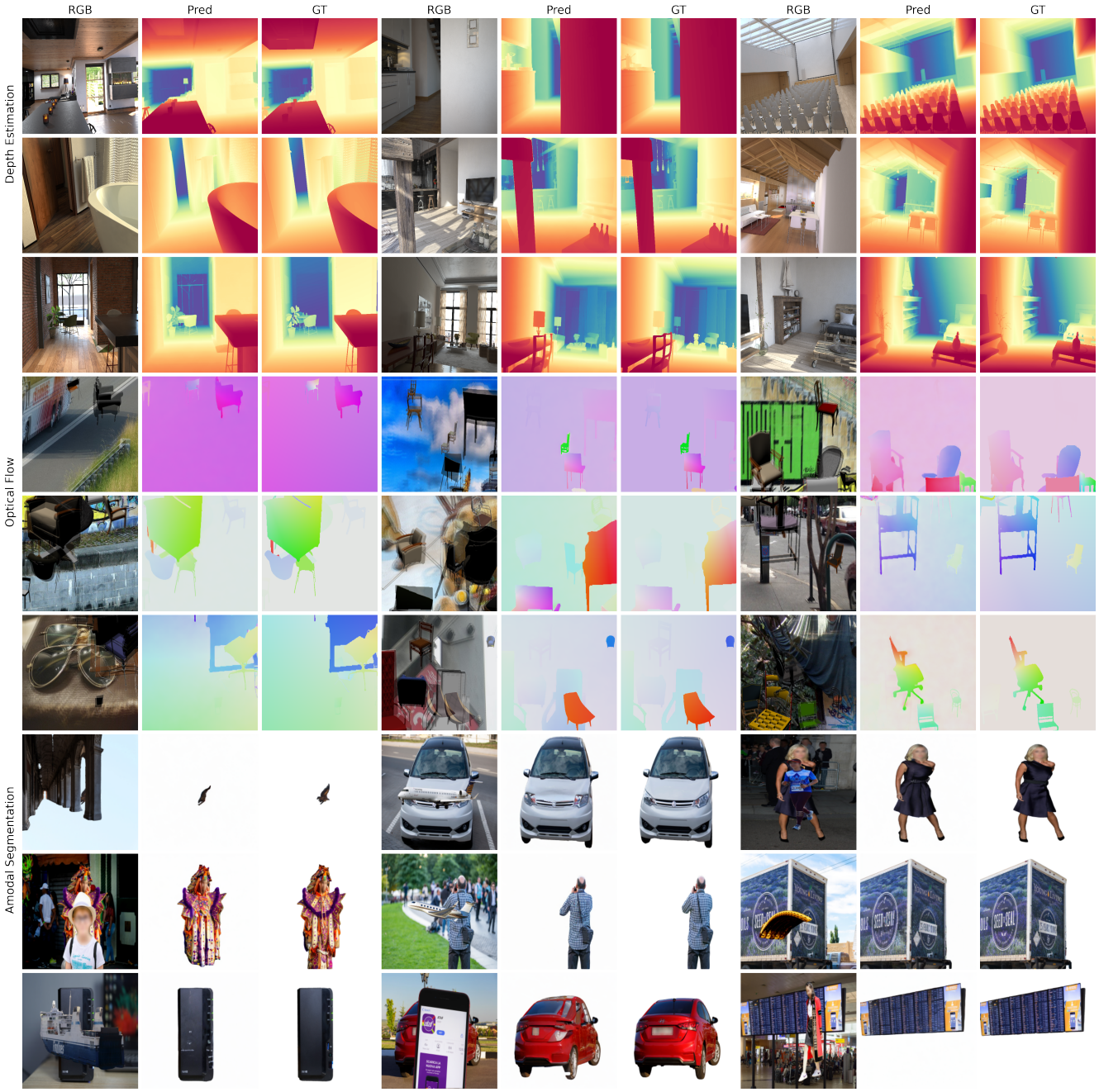

Figure 2. **Generalist Model Predictions**: We visualize additional samples generated from our mixture-of-experts generalist diffusion model. We generate the depth estimation samples from Hypersim, the optical flow estimation samples from FlyingChairs, and the amodal segmentation samples from Pix2Gestalt.