# Supplementary Material for Test-Time Prompt Tuning for Zero-Shot Depth Completion

In this supplementary material, we provide additional visual prompts of our approach. We first provide analysis of pixel-wise visual prompt in Sec. 1. We then provide analysis of variations of visual prompt design in Sec. 2.

## 1. Pixel-Wise Visual Prompts

In this section, we visualize the learned visual prompts for model input and the final input images that incorporate these visual prompts in Fig. 1. From a microscopic perspective, the visual prompts exhibit grid-like patterns, where the specific shapes within units correspond to the patch size of the transformer in the depth foundation model. Due to the self-attention mechanism in the transformer layer, these prompts are learned to have a structured formation. From a macroscopic perspective, the visual prompts exhibit distinct patterns for different objects identified in the image. In particular, they are trained to adopt varying patterns to clearly distinguish differences in distance between two wall planes. This structured prompting aids the depth foundation model in accurately predicting depth scales.

## 2. Various Prompt Design

### 2.1. Square Lattice

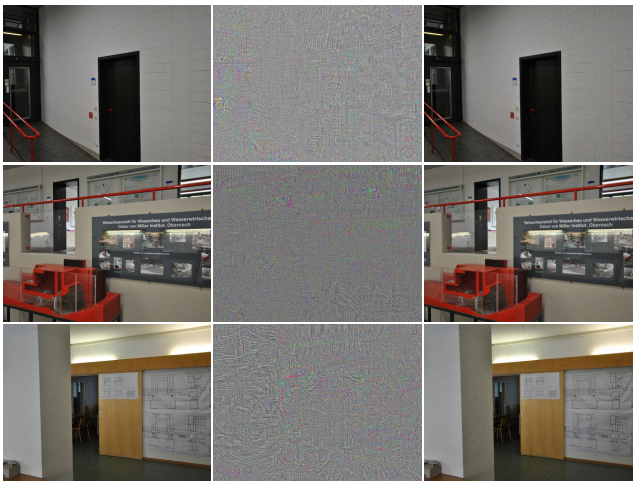
In this section, we present various strategies for visual prompt design. Firstly, we design a visual prompt by training only square lattice pixels at varying distances, rather than training all pixels. As shown in Tab. 1, the performance differences are marginal. However, we observe slightly better performance at 6% and 3% sparsity levels. While this approach does not significantly reduce memory usage, it demonstrates the possibilities that training a structured subset of pixels can be more effective than utilizing all pixels.

### 2.2. Random Pixel Selection

Next, instead of training all pixels, we randomly select learnable pixels across all visual prompts. As shown in Tab. 1, this random sampling approach achieves similar but slightly better performance compared to the square lattice method. This suggests that visual prompting can be more effective when not restricted to fixed positions.

### 2.3. Sparse Patch Selection

Lastly, we evaluate the sparse patch selection approach for visual prompts. Inspired by the observation that visual prompts naturally exhibit patch-like structures, we design a patch-level selection strategy. Specifically, we select learnable patch regions only when the area contains sparse depth information. As shown in Tab. 2, the sparse patch-based prompting approach leads to further improvements compared to the random selection method. These findings suggest that visual prompts can be learned more effectively when aligned with the positions of sparse depth data. Additionally, this opens up future research directions for discovering optimal strategies in pixel-wise prompt design.



(a) RGB (b) Visual Prompt (c) RGB + Visual Prompt

Figure 1. Visualization of the RGB image, the pixel-wise visual prompt, and their combined image.

| Method | 25% | | 6% | | 3% | | 2% | |
|---|---|---|---|---|---|---|---|---|
| | MAE↓ | RMSE↓ | MAE↓ | RMSE↓ | MAE↓ | RMSE↓ | MAE↓ | RMSE↓ |
| Squaer Lattice | 0.043 | 0.158 | 0.044 | 0.154 | 0.047 | 0.153 | 0.052 | 0.157 |
| Random Pixel | 0.043 | 0.159 | 0.044 | 0.153 | 0.046 | 0.152 | 0.049 | 0.154 |

Table 1. Comparison of different sampling strategies evaluated on the Ibims-1 dataset at sparsity levels of 25%, 6%, 3%, and 2%.

| Method | MAE↓ | RMSE↓ |
|---|---|---|
| Random Pixel Selection | 0.0430 | 0.1605 |
| Sparse Patch Selection | 0.0428 | 0.1589 |

Table 2. Comparison between random pixel and sparse patch selection strategies evaluated on the Ibims-1 dataset at sparsity levels of 25%, 6%, 3%, and 2%.