DICEPTION: A Generalist Diffusion Model for Visual Perceptual Tasks

Canyu Zhao, Yanlong Sun, Mingyu Liu, Huanyi Zheng,
Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, Chunhua Shen

Vision Still Lacks a One-For-All Model

Unlike NLP, computer vision suffers from a representational chasm. Foundation models remain monomaniacs—excelling at one task but failing to transfer without massive redesign.

SAM (Segment Anything)
Segments brilliantly, but cannot predict depth.

DepthAnything
Excels at depth, but cannot segment.

The Central Question: Can a single, compact fully parameter-sharing model deliver competitive accuracy across diverse visual understanding tasks?

The Potential of Diffusion Models in Multiple Perceptual Tasks

Powerful Generative Priors

Trained on billions of images, they learn rich geometry, context, and fine details.

Repurposed for Understanding

By unifying the representations of different tasks into RGB space.

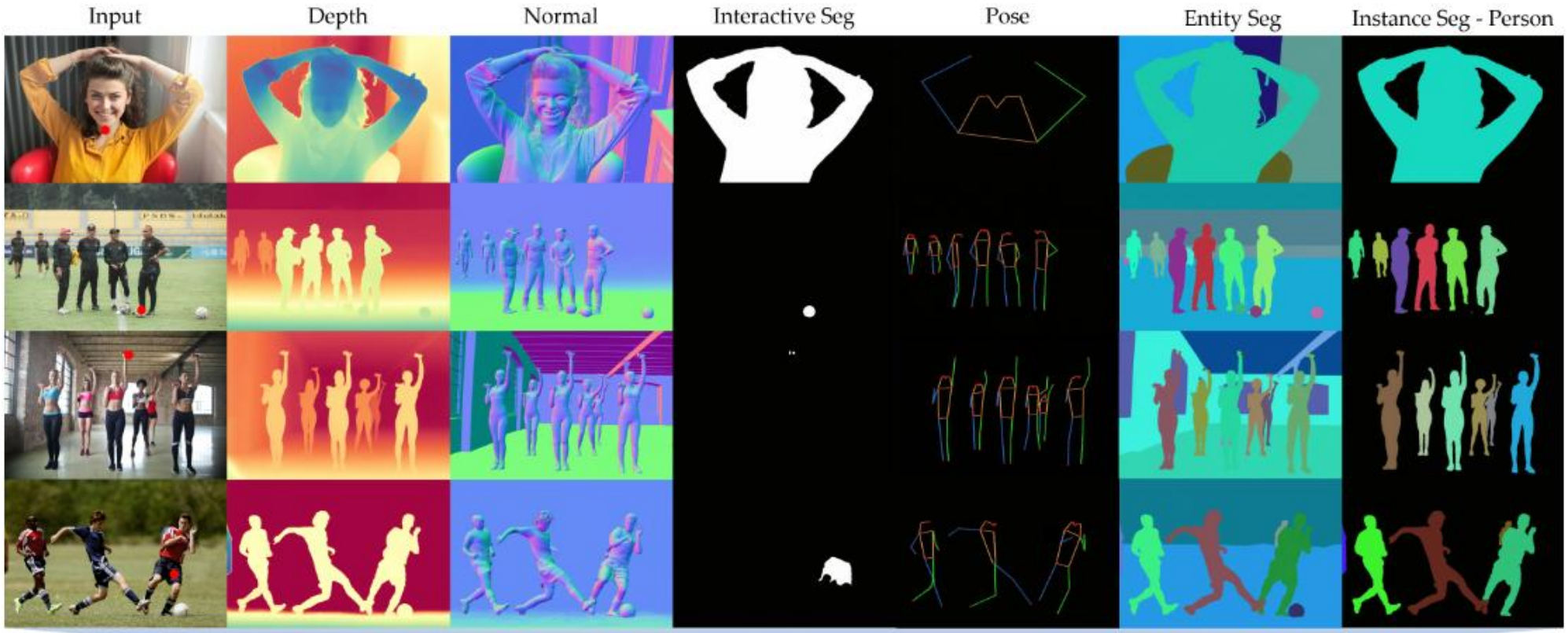
By repurposing pretrained diffusion weights, we can bypass the massive data and training costs of specialist models for multiple perception tasks.

DICEPTION

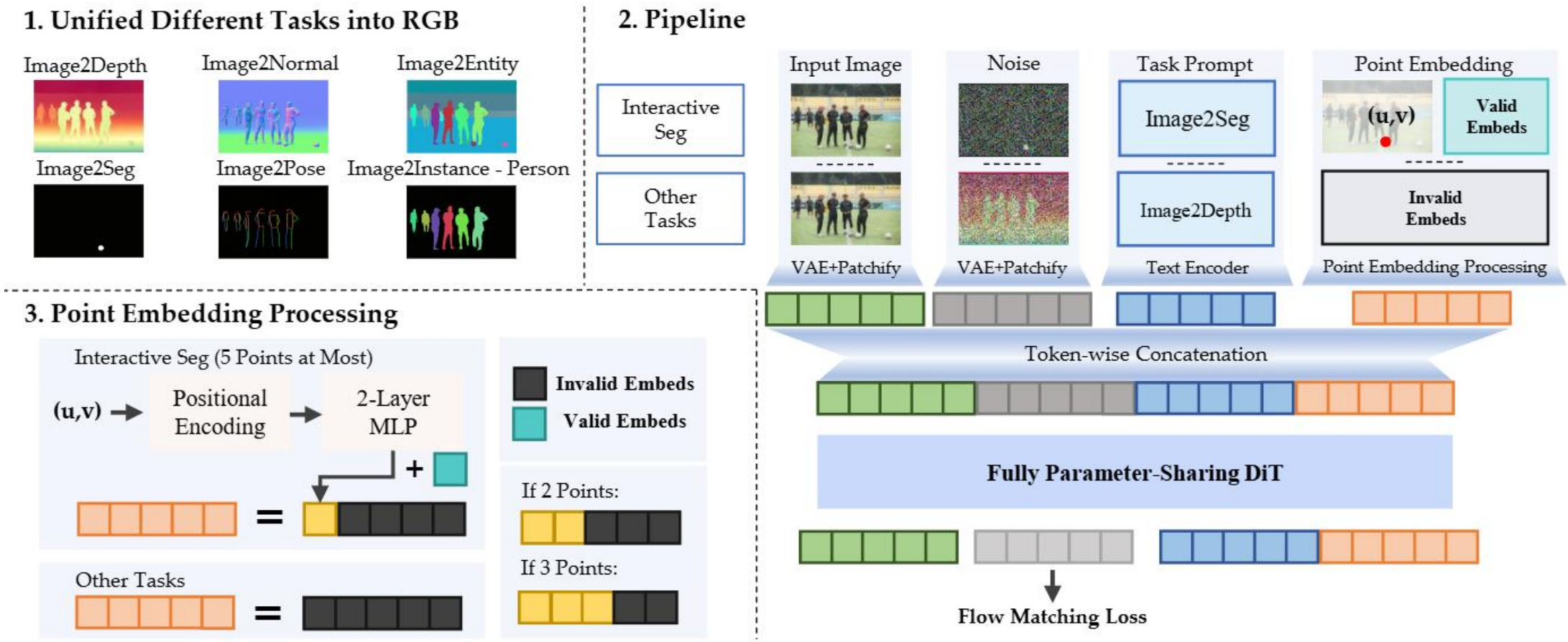
A single diffusion-based visual generalist model that handles **six perception tasks**—depth, normal, keypoint, and three types of segmentation—using **one shared set of parameters** and a unified RGB output space, all without task-specific modules. DICEPTION demonstrates strong scalability and adaptability.

What Can DICEPTION Do?

- **With one single model**, DICEPTION solves 6 perception tasks without relying on any task-specific modules, **on par with SOTA.**
- DICEPTION can quickly adapt to new tasks by fine-tuning less than **1%** of its parameters on as few as **50 images.**



Method


1. Unified Different Tasks into RGB
Image2Depth
Image2Normal
Image2Entity
Image2Seg
Image2Pose
Image2Instance - Person
2. Pipeline
Interactive Seg
Other Tasks
Input Image
Noise
Task Prompt
Point Embedding
Image2Seg
Image2Depth
(u,v)
Valid Embeds
Invalid Embeds
VAE+Patchify
VAE+Patchify
Text Encoder
Point Embedding Processing
Token-wise Concatenation
Fully Parameter-Sharing DiT
Flow Matching Loss
3. Point Embedding Processing
Interactive Seg (5 Points at Most)
(u,v)
Positional Encoding
2-Layer MLP
Other Tasks
Invalid Embeds
Valid Embeds
If 2 Points:
If 3 Points:

Results - Depth

Method	Training Samples	KITTI [33]		NYUv2 [77]		ScanNet [24]		DIODE [106]		ETH3D [95]	
		AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
MiDaS [89]	2M	0.236	0.630	0.111	0.885	0.121	0.846	0.332	0.715	0.184	0.752
Omnidata [28]	12.2M	0.149	0.835	0.074	0.945	0.075	0.936	0.339	0.742	0.166	0.778
DPT-large [88]	1.4M	0.100	0.901	0.098	0.903	0.082	0.934	0.182	0.758	0.078	0.946
DepthAnything† [123]	63.5M	0.080	0.946	0.043	0.980	0.043	0.981	0.261	0.759	0.058	**0.984**
DepthAnything v2† [124]	62.6M	0.080	0.943	0.043	0.979	0.042	0.979	0.321	0.758	0.066	0.983
Depth Pro† [7]	-	0.055	0.974	0.042	0.977	0.041	0.978	0.217	0.764	0.043	0.974
Metric3D v2† [45]	16M	**0.052**	**0.979**	**0.039**	**0.979**	**0.023**	**0.989**	**0.147**	**0.892**	**0.040**	0.983
DiverseDepth [129]	320K	0.190	0.704	0.117	0.875	0.109	0.882	0.376	0.631	0.228	0.694
LeReS [130]	354K	0.149	0.784	0.090	0.916	0.091	0.917	0.271	0.766	0.171	0.777
HDN [132]	300K	0.115	0.867	0.069	0.948	0.080	0.939	0.246	0.780	0.121	0.833
GeoWizard [32]	280K	0.097	0.921	0.052	0.966	0.061	0.953	0.297	0.792	0.064	0.961
DepthFM [34]	63K	0.083	0.934	0.065	0.956	-	-	0.225	0.800	-	-
Marigold† [49]	74K	0.099	0.916	0.055	0.964	0.064	0.951	0.308	0.773	0.065	0.960
DMP Official† [56]	-	0.240	0.622	0.109	0.891	0.146	0.814	0.361	0.706	0.128	0.857
GeoWizard† [32]	280K	0.129	0.851	0.059	0.959	0.066	0.953	0.328	0.753	0.077	0.940
DepthFM† [34]	63K	0.174	0.718	0.082	0.932	0.095	0.903	0.334	0.729	0.101	0.902
Genpercept† [118]	90K	0.094	0.923	0.091	0.932	0.056	0.965	0.302	0.767	0.066	0.957
Painter† [113]	24K	0.324	0.393	**0.046**	**0.979**	0.083	0.927	0.342	0.534	0.203	0.644
Unified-IO† [71]	48K	0.188	0.699	0.059	0.970	**0.063**	**0.965**	0.369	0.708	0.103	0.906
4M-XL† [75]	759M	0.105	0.896	0.068	0.951	0.065	0.955	0.331	**0.734**	0.070	0.953
OneDiffusion† [55]	500K	0.101	0.908	0.087	0.924	0.094	0.906	0.399	0.661	0.072	0.949
Ours-single†	500K	0.064	0.952	0.066	0.953	0.077	0.942	0.283	0.717	0.052	0.971
Ours†	500K	**0.069**	**0.949**	0.061	0.960	0.072	0.944	**0.289**	0.722	**0.050**	**0.975**

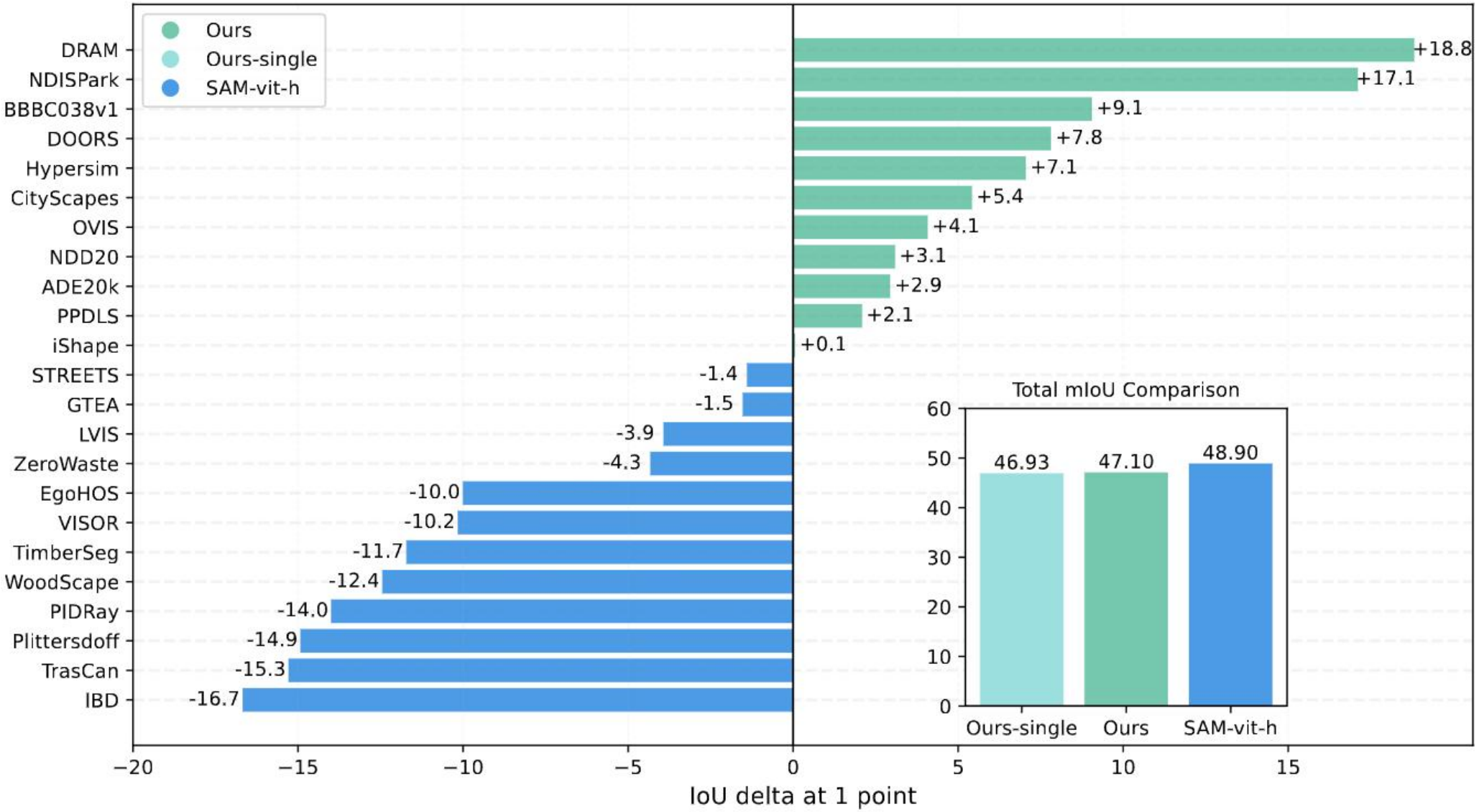
Results - Normal

Method	Training Samples	NYUv2 [77]					ScanNet [24]					DIODE-indoor [106]				
		mean↓	med↓	11.25°↑	22.5°↑	30°↑	mean↓	med↓	11.25°↑	22.5°↑	30°↑	mean↓	med↓	11.25°↑	22.5°↑	30°↑
DINSE [3]	160K	**18.572**	10.845	**54.732**	74.146	80.256	18.610	9.885	56.132	76.944	82.606	18.453	13.871	36.274	77.527	86.976
Geowizard [32]	280K	20.363	11.898	46.954	73.787	80.804	19.748	9.702	58.427	77.616	81.575	19.371	15.408	30.551	75.426	86.357
GenPercept [118]	90K	20.896	11.516	50.712	73.037	79.216	18.600	8.293	64.697	79.329	82.978	18.348	13.367	39.178	79.819	88.551
Marigold [49]	90K	20.864	11.134	50.457	73.003	79.332	18.463	8.442	64.727	79.559	83.199	16.671	12.084	45.776	82.076	89.879
StableNormal [127]	250K	19.707	**10.527**	53.042	**75.889**	**81.723**	**17.248**	**8.057**	**66.655**	**81.134**	**84.632**	**13.701**	**9.460**	**63.447**	**86.309**	**92.107**
Unified-IO [70]	210K	28.547	14.637	39.907	63.912	71.240	**17.955**	10.269	**54.120**	**77.617**	**83.728**	31.576	16.615	27.855	64.973	73.445
4M-XL [75]	759M	37.278	13.661	44.660	60.553	65.327	30.700	11.614	48.743	68.867	73.623	18.189	12.979	36.622	81.844	87.050
Ours-single	500K	18.292	10.145	52.693	76.966	83.041	18.807	10.327	52.919	75.152	82.968	16.229	11.012	50.137	83.573	88.972
Ours	500K	**18.338**	**10.106**	**52.850**	**77.079**	**82.903**	18.842	**10.266**	53.610	74.895	82.864	**16.297**	**11.117**	**50.548**	**83.325**	**88.774**

Results - Interactive Segmentation

0.06%

Data vs. SAM



TakeAways

Token-wise vs. Channel-wise Concatenation

Token-wise

- Faster convergence
- Higher accuracy
- No extra parameters
- Better leverages pretrained weights

Channel-wise

- Slower convergence
- Lower accuracy
- Requires extra MLP
- Does not fully leverage pretrained weights

Token-wise fusion preserves the pretrained prior most effectively.

Architecture Matters: DiT vs. U-Net

A U-Net baseline fails at multitasking due to information loss from downsampling and disrupted priors from extra convolutional layers.

DiT's pure transformer maintains full-resolution pathways and a global receptive field, leading to markedly superior geometry predictions.

U-Net Baseline

Fail even on simple two-task cases.

U-Net Baseline

Achieves superior results without architectural redesign.

The Effect of Classifier-Free Guidance

A mild CFG scale improves dense tasks like depth and normal estimation but has little effect on sparse tasks like segmentation and human keypose estimation.

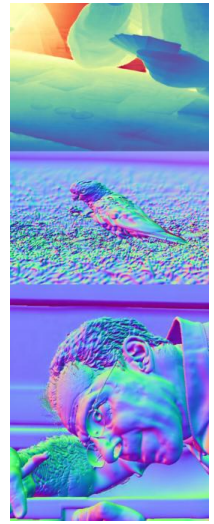
CFG = 1

Softer, less defined

CFG = 2 (Optimal)

Sharper, clearer details

CFG > 3

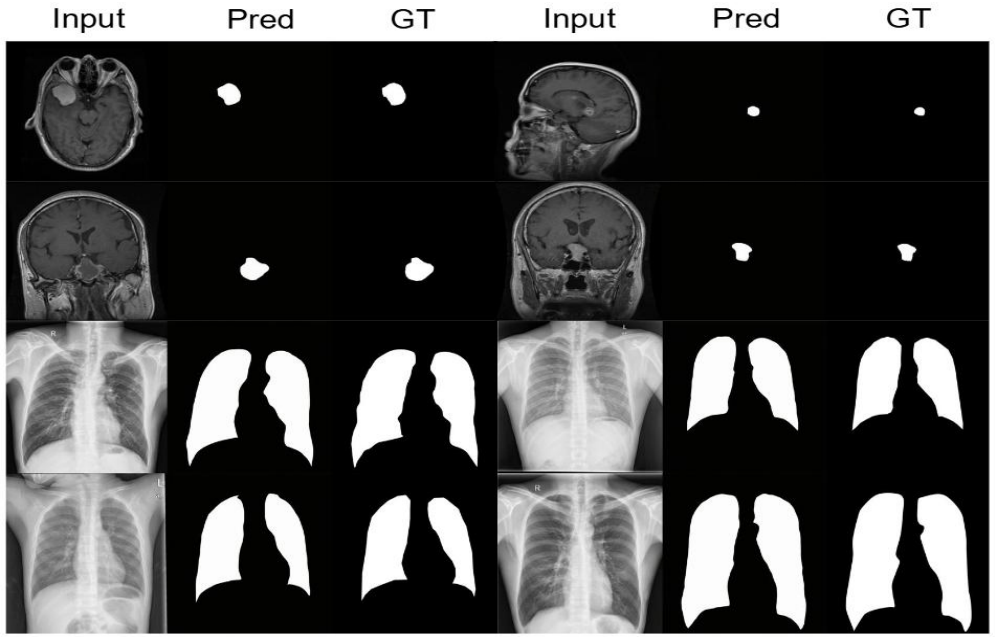


Over-saturated, artifacts

Few - Shot Adaptation

Rapid adaptation to new tasks with minimal data and trainable parameters.

50 Images + 1% Parameters



Efficiency: Few-Step Inference

Thanks to flow-matching's straight trajectory, DICEPTION supports few-step inference with minimal performance degradation on **Dense Tasks such as Depth and Normal.**

Ours (1-Step)

Depth AbsRel: 0.086

OneDiffusion (1-Step)

Depth AbsRel: FAIL

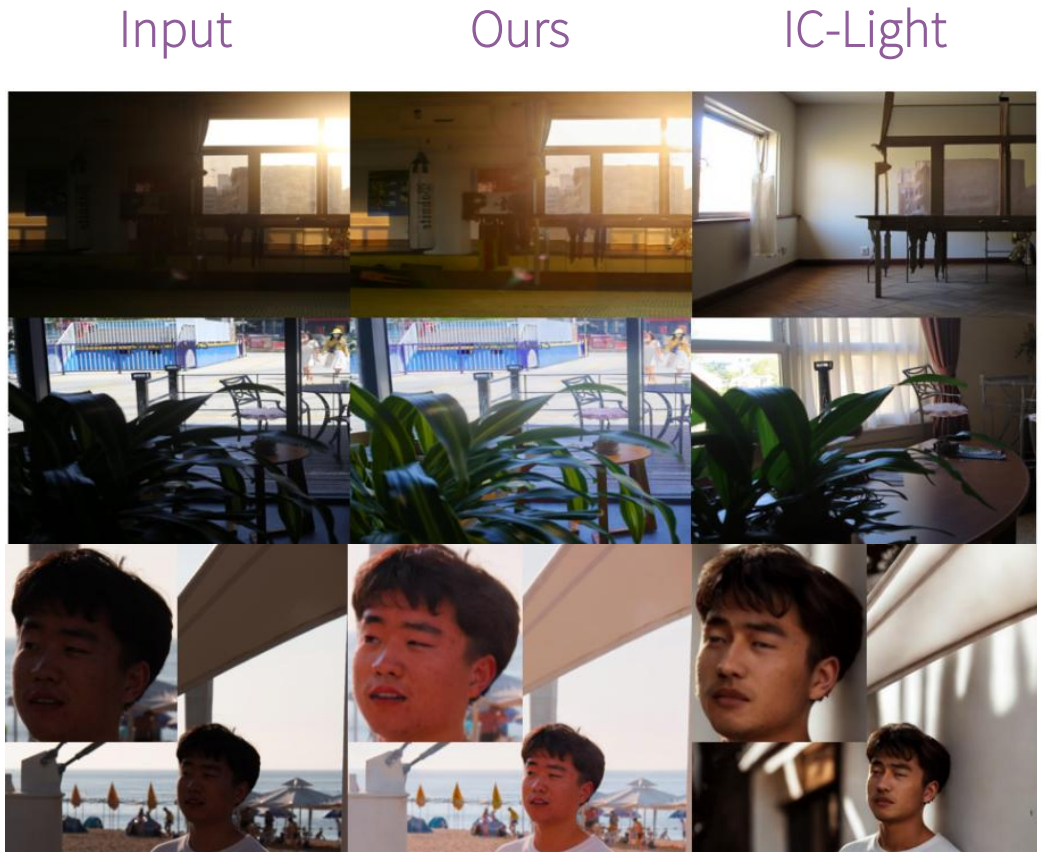
This built-in acceleration opens a realistic path to real-time deployment without extra distillation.

Method	Training Samples	KITTI [33]		NYUv2 [77]		ScanNet [24]		DIODE [106]		ETH3D [95]	
		AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
28-step	500K	0.069	0.949	0.061	0.960	0.072	0.944	0.289	0.722	0.050	0.975
14-step	500K	0.077	0.942	0.063	0.958	0.074	0.943	0.272	0.718	0.048	0.978
7-step	500K	0.081	0.939	0.065	0.953	0.078	0.943	0.286	0.714	0.052	0.971
3-step	500K	0.083	0.938	0.069	0.953	0.077	0.940	0.294	0.707	0.063	0.967
1-step	500K	0.086	0.936	0.072	0.945	0.076	0.937	0.305	0.702	0.065	0.967

Pixel-Aligned Training

Our perceptual settings enforce strict pixel correspondence, which significantly reduces generative artifacts in downstream applications.

It is important to emphasize that our goal is not to compare the lighting quality between methods, but rather to highlight our model's ability to significantly reduce generative artifacts and retain structural details.



Conclusion & Roadmap

Key Takeaways

- Diffusion priors can power strong multitask visual understanding.
- Five critical designs enable this: Token fusion, DiT, RGB unification, light CFG, and pixel-aligned training.
- Achieves SOTA-comparable performance without massive curated data.

Future Roadmap

- Scale data and model size for even better performance.
- Further compress inference steps for real-time use.
- Expand to more tasks for a truly universal visual foundation model.

Thank you

DICEPTION: A Generalist Diffusion Model for Visual Perceptual Tasks

Canyu Zhao, Yanlong Sun, Mingyu Liu, Huanyi Zheng,
Muzhi Zhu, Zhiyue Zhao, Hao Chen, Tong He, Chunhua Shen