TransNormal: Dense Visual Semantics for Diffusion-based Transparent Object Normal Estimation

Anonymous Authors[1]

Abstract

Monocular normal estimation for transparent objects is critical for laboratory automation, yet it remains challenging due to complex light refraction and reflection. These optical properties often lead to catastrophic failures in conventional depth and normal sensors, hindering the deployment of embodied AI in scientific environments. We propose **TransNormal**, a novel framework that adapts pre-trained diffusion priors for single-step normal regression. To handle the lack of texture in transparent surfaces, TransNormal integrates dense visual semantics from DINOv3 via a cross-attention mechanism, providing strong geometric cues. Furthermore, we employ a multi-task learning objective and wavelet-based regularization to ensure the preservation of fine-grained structural details. To support this task, we introduce **TransNormal-Synthetic**, a physics-based dataset with high-fidelity normal maps for transparent labware. Extensive experiments demonstrate that TransNormal significantly outperforms state-of-the-art methods: on the ClearGrasp benchmark, it reduces mean error by 24.4% and improves 11.25° accuracy by 22.8%; on ClearPose, it achieves a 15.2% reduction in mean error. The code and dataset will be made publicly available.

1. Introduction

Motivation. Embodied AI agents hold the potential to significantly accelerate scientific discovery in autonomous laboratory environments (Tao et al., 2025; Fang et al., 2023; Wen et al., 2024). However, a primary barrier to their practical deployment is the perceptual instability caused by variable illumination (Tobin et al., 2017; James et al., 2019; Lind et al., 2024): variations in lighting and shadows can induce significant appearance shifts, leading to unstable detection and degraded manipulation performance. Unlike intensity-based features, surface normal maps provide a lighting-invariant geometric representation, offering a stable cue for perception and manipulation in real-world laboratories where lighting is often uncontrolled.

Despite the maturity of normal estimation for opaque objects, transparent labware—such as beakers, pipettes, and culture dishes—presents a unique and formidable challenge. The difficulty of estimating normals for these objects is three-fold: ① **Geometrically**, transparent surfaces often lack discriminative textures and exhibit indistinct boundaries. Furthermore, multi-layered interfaces (*e.g.*, glass-air-liquid) introduce significant structural ambiguities that are absent in opaque surfaces. ② **Optically**, the dominance of refraction and reflection makes the visual appearance of labware highly dependent on the surrounding environment, often rendering the objects nearly invisible to standard sensors. ③ **Perceptually**, accurate reconstruction requires high-level reasoning, including object-level shape priors (*e.g.*, the canonical geometry of a beaker) and contextual inference to distinguish between transparent and opaque material regions. Consequently, traditional geometric cues such as shading, texture gradients, and edge detection, which are the cornerstones of normal estimation for opaque objects, become unreliable or entirely absent. This necessitates a more robust approach that can leverage deep priors to resolve the inherent ambiguities of transparent surfaces.

Given the physical complexities of light transport, monocular normal estimation for transparent objects necessitates high-level reasoning about materials and global shapes. However, most existing frameworks (Bae & Davison, 2024) predominantly treat this as a localized regression task, relying on local image or photometric cues. While effective for textured opaque surfaces, these inductive biases are fundamentally ill-suited for transparent labware, where refraction and reflection decouple local appearance from underlying geometry. Furthermore, current research in the transparent domain has focused largely on 3D shape estimation, depth completion or 6D pose estimation (Sajjan et al., 2020; Chen et al., 2022; Fang et al., 2022; Kim et al., 2024), leaving

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

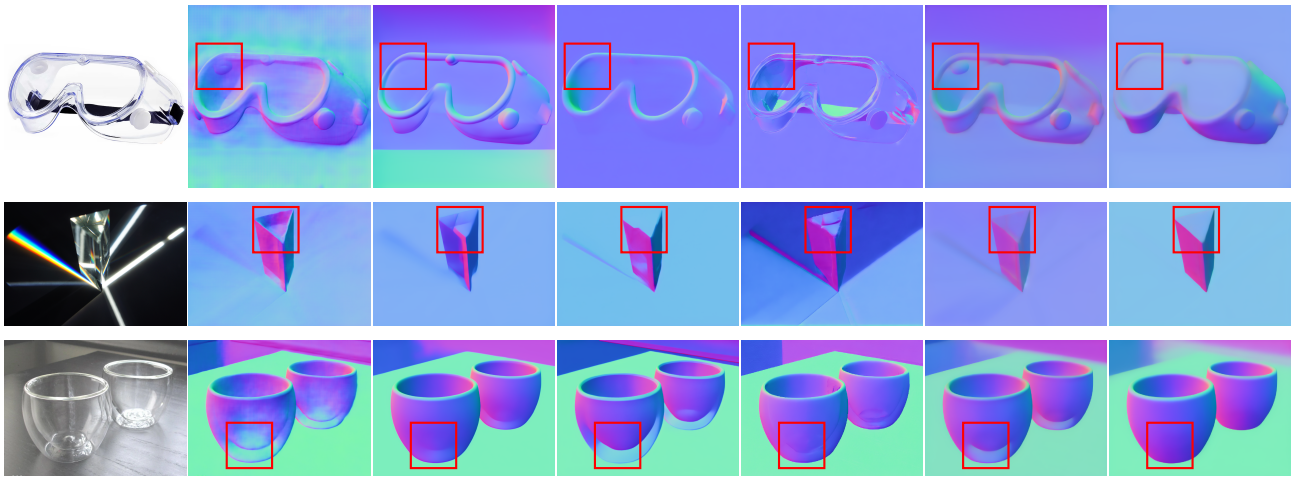


Figure 1. **In-the-wild qualitative results: TransNormal achieves accurate surface normal estimation for transparent objects. Row 1 (Safety goggles):** The vent holes in the upper-left corner lie behind the lens and should not affect surface normals; the lens itself should appear flat. Only our method correctly recovers the flat lens without being misled by background textures visible through refraction. **Row 2 (Prism):** Under extreme lighting with complex refraction and caustics, only our method recovers the correct triangular geometry without being distracted by background colored light. **Row 3 (Double-walled glass):** Other methods produce artifacts from the inner glass layer, while ours correctly estimates the outer surface normals.

the estimation of dense surface normals, which is a critical representation for fine-grained robotic manipulation and liquid handling, relatively under-explored. The lack of high-quality benchmarks with dense normal annotations further hinders progress.

Our key insight is that the inherent ambiguities of transparent surfaces can be resolved by leveraging high-level scene understanding and physical priors encoded in large-scale vision models. Unlike local discriminative kernels, generative models trained on diverse web-scale data may already capture the "canonical" geometry and material properties of objects. This motivates the use of model families whose conditioning pathways allow for task-specific guidance to bridge the gap between low-level appearance and high-level geometric structure. Diffusion-based dense prediction provides such a model family. Recent advances (Ke et al., 2024; Fu et al., 2024) demonstrate that pre-trained text-to-image models, such as Stable Diffusion, possess rich geometric and material priors. However, we observe a significant gap in current practice: many methods (He et al., 2025; Ke et al., 2024; Zhao et al., 2025) utilize these models with empty or generic text prompts, leaving the cross-attention conditioning pathway, which is originally designed for complex semantic alignment, largely underutilized for geometric tasks.

We propose **TransNormal**, a framework that repurposes Stable Diffusion's conditioning mechanism for dense semantic injection. Rather than relying on sparse text, we inject dense visual semantics from DINOv3 (Siméoni et al., 2025) into the diffusion backbone. By transforming cross-attention into a semantic-geometric guidance channel, TransNormal effectively resolves the geometric ambiguities of transparent labware using global context. To facilitate robust training and evaluation, we introduce TransNormal-Synthetic, a physics-based dataset with high-fidelity normal maps for transparent labware. Despite being trained on only ∼122K synthetic samples, TransNormal achieves state-of-the-art performance on ClearGrasp (Sajjan et al., 2020) and ClearPose (Chen et al., 2022) (Tab. 1), reducing mean error on real-world data by significant margins. Our key contributions are as follows:

- **Semantic-Geometric Conditioning:** We identify a critical underutilization in diffusion-based dense prediction and propose to replace sparse text conditioning with dense DINOv3 visual semantics to provide material-aware geometric guidance.
- **TransNormal Framework:** We present a novel architecture that adapts Stable Diffusion for single-step normal regression. TransNormal achieves superior generalization to transparent surfaces with significantly fewer training samples than traditional Transformer-based discriminative baselines.
- **Physics-Based Dataset:** We introduce TransNormal-Synthetic, a high-quality benchmark providing physically accurate normal maps rendered from 3D labware meshes, enabling controlled and systematic evaluation of transparent object perception.
- **State-of-the-Art Performance:** Our method sets new performance standards across multiple benchmarks.

On ClearGrasp, TransNormal reduces mean angular error by 24.4% and improves 11.25° accuracy by 22.8%; on the real-world ClearPose dataset, it achieves a 15.2% error reduction. These results demonstrate robust zero-shot transfer from synthetic training to complex, real-world laboratory environments.

2. Related Work

2.1. Geometric Dense Prediction and Generative Priors

Recovering geometric properties such as depth and surface normals has evolved through three paradigms. Early physics-based methods relied on Structure from Motion (SfM) (Tomasi & Kanade, 1992), photometric stereo (Woodham, 1980), and multi-view geometry (Scharstein & Szeliski, 2002), but were brittle under real-world conditions. The discriminative learning paradigm (Eigen et al., 2014; Eftekhar et al., 2021; Ranftl et al., 2020) and recent large-scale models like MoGe (Wang et al., 2025a;b) and Depth Anything (Yang et al., 2024a;b) achieved remarkable success, yet struggle with out-of-distribution scenarios such as transparent or reflective surfaces.

Most recently, a **generative paradigm** has emerged, reframing dense prediction as conditional generation. Models like Marigold (Ke et al., 2024) and GeoWizard (Fu et al., 2024) leverage world priors from large-scale diffusion models (Rombach et al., 2022) for strong zero-shot generalization. The adaptation of these priors follows three trajectories: (a) *stochastic generative* methods (*e.g.*, Marigold, DepthFM (Gui et al., 2025)) use multi-step diffusion but suffer from inference inefficiency and structural variance; (b) *deterministic feed-forward* approaches (*e.g.*, DiffusionE2E-FT (Martin Garcia et al., 2025), Lotus (He et al., 2025)) fine-tune backbones for speed but often lose fine-grained details; (c) *coarse-to-fine* strategies (*e.g.*, StableNormal (Ye et al., 2024)) bridge this gap but often reintroduce stochasticity in refinement. A critical limitation across these methods is their underutilization of semantic conditioning—they typically use empty text prompts or simple category labels, leaving rich semantic priors largely unexploited. This overlooks the potential of dense visual semantics: recent self-supervised encoders like DINOv2 (Oquab et al., 2023) and DINOv3 (Siméoni et al., 2025) capture robust object-centric representations that persist even under refractive distortions, offering a more suitable guidance signal for geometry estimation. Our work builds upon this generative paradigm, integrating such dense semantic guidance to address the challenges of transparent materials.

2.2. Geometry Estimation for Transparent Objects

Perception of transparent objects is uniquely challenging due to refraction and reflections that cause commodity depth sensors to produce large holes or distortions. Related tasks include transparent object segmentation (Xie et al., 2020; 2021; Sun et al., 2023) and 6D pose estimation (Zhang et al., 2022; Jiang et al., 2024), which share similar optical challenges. Early methods like ClearGrasp (Sajjan et al., 2020) and DREDS (Dai et al., 2022) pioneered learning-based depth completion, while subsequent work (Zhu et al., 2021; Xu et al., 2022; Hong et al., 2022; Cai et al., 2023) recovered true depth from corrupted RGB-D inputs, enabled by benchmarks like ClearPose (Chen et al., 2022) and TransCG (Fang et al., 2022). Physics-based approaches have also explored monocular shape from refraction (Sulc et al., 2021) and refractive flow for normal estimation (Tang et al., 2024), while polarization cameras offer complementary normal cues (Shao et al., 2023). When multi-view RGB is accessible, neural implicit representations enable full geometry recovery (Ichnowski et al., 2021; Li et al., 2023; Zhou et al., 2023; Deng et al., 2024; Sun et al., 2024; Li et al., 2025), though requiring dense viewpoints. Very recently, video diffusion models (Hu et al., 2025) have been adapted for geometry estimation, with DKT (Xu et al., 2025b) extending this paradigm to transparent object depth; however, these methods require temporal sequences as input, limiting their applicability to single-image scenarios. Accurate geometry also underpins robotic manipulation (Bai et al., 2023; 2025b;a). Training data has evolved from Physics-Based Rendering (PBR) (Sajjan et al., 2020) to generative synthesis (Zhang & Agrawala, 2024; Agrawal et al., 2024). Our approach fine-tunes a generative backbone on a curated synthetic dataset that disentangles geometry from material appearance, internalizing a rich prior of transparent phenomena.

3. Preliminaries

Latent Diffusion Models. Our framework is built upon Stable Diffusion (Rombach et al., 2022), which performs the diffusion process in a compressed latent space for computational efficiency. This is enabled by a pre-trained Variational Auto-Encoder (VAE) consisting of an encoder $\mathcal{E}(\cdot)$ and a decoder $\mathcal{D}(\cdot)$, which maps between RGB space and latent space, *i.e.*, $\mathcal{E}(\boldsymbol{x}) = \boldsymbol{z}^x$, $\mathcal{D}(\boldsymbol{z}^x) \approx \boldsymbol{x}$. Following recent dense prediction works (Ke et al., 2024; Fu et al., 2024; Xu et al., 2025a; Ye et al., 2024), we also map dense annotations into this latent space: $\mathcal{E}(\boldsymbol{y}) = \boldsymbol{z}^y$, $\mathcal{D}(\boldsymbol{z}^y) \approx \boldsymbol{y}$.

Diffusion Process. Stable Diffusion establishes a probabilistic model through a *forward* noising process and a *reversal* denoising process. In the *forward* process, Gaussian noise is gradually added to the latent $\boldsymbol{z}^y$ over time steps $t \in [1, T]$ to obtain the noisy sample $\boldsymbol{z}_t^y = \sqrt{\overline{\alpha}_t}\boldsymbol{z}^y + \sqrt{1-\overline{\alpha}_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$, $\overline{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$, and $\{\beta_1, \beta_2, \ldots, \beta_T\}$ is the noise schedule. At time-step T, the sample $\boldsymbol{z}_T^y$ approximates pure Gaussian noise. In the *reversal* process, a

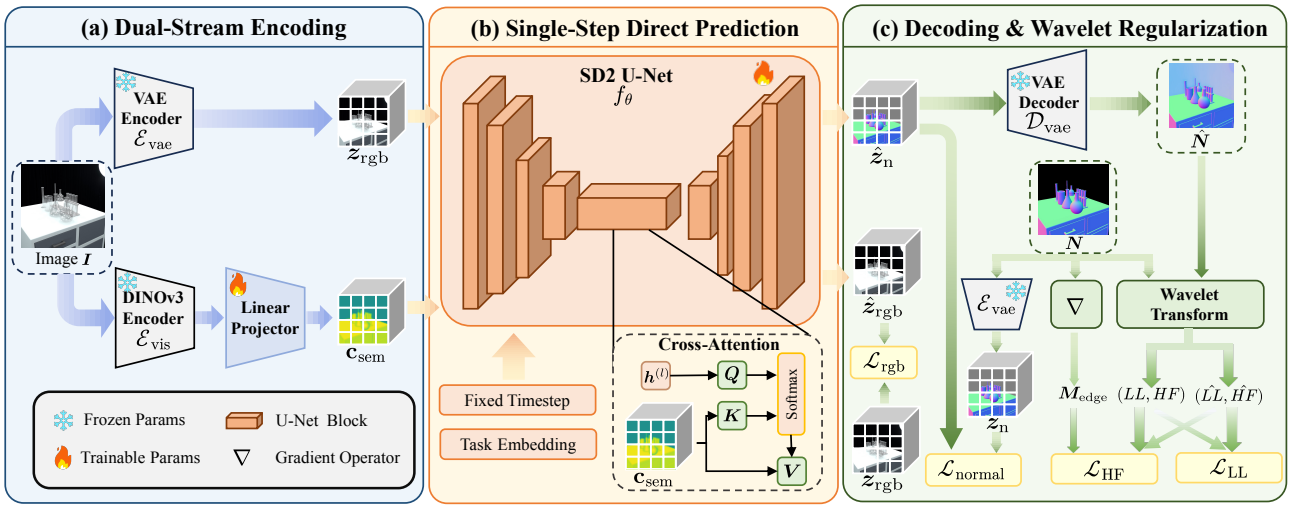


Figure 2. **Overview of the TransNormal framework.** (a) **Dual-Stream Encoding**: the frozen VAE encoder $\mathcal{E}_{\text{vae}}$ extracts RGB latent $\boldsymbol{z}_{\text{rgb}}$, while a frozen DINOv3 encoder $\mathcal{E}_{\text{vis}}$ with a trainable linear projector produces semantic conditioning $\boldsymbol{c}_{\text{sem}}$; (b) **Single-Step Direct Prediction**: the fine-tuned SD2 U-Net $f_{\boldsymbol{\theta}}$ directly regresses the normal latent $\hat{\boldsymbol{z}}_{\text{n}}$ from $\boldsymbol{z}_{\text{rgb}}$ at a fixed timestep T, where the spatial feature $\boldsymbol{h}^{(l)}$ at each layer l provides queries $\boldsymbol{Q}$, and $\boldsymbol{c}_{\text{sem}}$ provides keys $\boldsymbol{K}$ and values $\boldsymbol{V}$ for cross-attention; (c) **Decoding & Wavelet Regularization**: the frozen VAE decoder $\mathcal{D}_{\text{vae}}$ reconstructs the predicted normal map $\hat{\boldsymbol{N}}$, supervised by latent-space losses ($\mathcal{L}_{\text{rgb}}$, $\mathcal{L}_{\text{normal}}$) and wavelet-domain losses ($\mathcal{L}_{\text{HF}}$, $\mathcal{L}_{\text{LL}}$) that separately penalize high-frequency details and low-frequency structure. (§ 4)

U-Net $f_{\boldsymbol{\theta}}$ (Ronneberger et al., 2015) is trained to iteratively remove noise from $\boldsymbol{z}_t^y$ to recover the clean sample $\boldsymbol{z}^y$.

Single-Step Regression for Dense Prediction. While the standard diffusion formulation relies on iterative sampling for stochastic generation, dense prediction tasks (*e.g.*, normal estimation) are inherently deterministic. Recent studies (Ke et al., 2024; He et al., 2025; Xu et al., 2025a) demonstrate that the pre-trained U-Net can be effectively repurposed for direct regression. Adopting this strategy, we simplify the inference process: instead of multi-step denoising, we fix the timestep at T and train the network to directly predict the clean annotation latent $\boldsymbol{z}^y$ from the input image latent $\boldsymbol{z}^x$ in a single forward pass:

$$\hat{\boldsymbol{z}}^y = f_{\boldsymbol{\theta}}(\boldsymbol{z}^x, T). \tag{1}$$

This approach leverages the strong priors of Stable Diffusion while ensuring deterministic and efficient prediction.

Notation. For clarity in the subsequent method description (§ 4), we adopt more descriptive subscripts: $\boldsymbol{z}^x \equiv \boldsymbol{z}_{\text{rgb}}$ denotes the RGB image latent and $\boldsymbol{z}^y \equiv \boldsymbol{z}_{\text{n}}$ denotes the normal map latent.

4. Method

Method Overview. Given an input RGB image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, our goal is to predict the surface normal map $\boldsymbol{N} \in \mathbb{R}^{H\times W\times 3}$. We build **TransNormal** by repurposing Stable Diffusion 2 (SD2) as a single-step normal predictor with semantic conditioning. This section follows SD2's data flow: (a) encoders and semantic conditioning; (b) the U-Net prediction module; and (c) VAE decoding and training objectives.

4.1. Dual-Stream Encoding

Semantic Guidance via Visual Prompting. Previous diffusion-based methods often rely on CLIP (Radford et al., 2021) text encoders with generic or empty prompts, leaving the powerful cross-attention mechanism underutilized. For transparent objects, where refraction corrupts local textures, such sparse conditioning is insufficient. We therefore replace the text encoder with a frozen DINOv3 visual encoder $\mathcal{E}_{\text{vis}}$ to extract dense, object-level semantic features:

$$\boldsymbol{F}_{\text{sem}} = \mathcal{E}_{\text{vis}}(\boldsymbol{I}) \in \mathbb{R}^{N_p \times d_{\text{dino}}}, \tag{2}$$

where $N_p = \lfloor H/p \rfloor \times \lfloor W/p \rfloor$ is the number of patch tokens with patch size p and feature dimension d_{dino}. These features are then projected into the U-Net's cross-attention dimension via a trainable linear projector $\boldsymbol{W}_{\text{proj}} \in \mathbb{R}^{d_{\text{dino}} \times d_{\text{unet}}}$:

$$\boldsymbol{c}_{\text{sem}} = \boldsymbol{F}_{\text{sem}} \boldsymbol{W}_{\text{proj}} \in \mathbb{R}^{N_p \times d_{\text{unet}}}. \tag{3}$$

This stream effectively acts as a dense "visual prompt", injecting robust semantic priors that persist even under refractive distortions. The DINOv3 encoder is kept frozen and only the lightweight projector $\boldsymbol{W}_{\text{proj}}$ is trained.

Latent Content Encoding. To leverage the generative priors of Stable Diffusion, the second stream maps the input image into the model's native latent space using the frozen VAE encoder $\mathcal{E}_{\text{vae}}$:

$$\boldsymbol{z}_{\text{rgb}} = \mathcal{E}_{\text{vae}}(\boldsymbol{I}) \in \mathbb{R}^{h\times w\times 4}, \tag{4}$$

where $(h, w) = (\lfloor H/8 \rfloor, \lfloor W/8 \rfloor)$. This latent representation $\boldsymbol{z}_{\text{rgb}}$ serves as the direct input to the U-Net, preserving spatial structure and fine-grained details for the regression task. Similarly, during training, the ground truth normal map $\boldsymbol{N}$ is also encoded into the latent space:

$$\boldsymbol{z}_{\text{n}} = \mathcal{E}_{\text{vae}}(\boldsymbol{N}) \in \mathbb{R}^{h \times w \times 4}. \tag{5}$$

4.2. Single-Step Prediction with Semantic Injection

Detail Preserver via Dual-Task Learning. To avoid catastrophic forgetting when fine-tuning a pre-trained diffusion model (Zhai et al., 2023), we follow He et al. (2025) and *adopt* their task switcher with two *fixed* task embeddings $s \in \{s_{\text{n}}, s_{\text{rgb}}\}$, added to the time embedding as class-label conditions. The same U-Net $f_{\boldsymbol{\theta}}$ serves both tasks: s_{n} triggers normal prediction and s_{rgb} triggers RGB reconstruction. These embeddings are kept fixed during training. This switch preserves fine detail while adapting the model to geometry.

Single-Step Normal Prediction. Unlike prior diffusion-based methods that inject noise and recover clean latents through iterative denoising, we directly input clean RGB latents and predict normal latents in a single forward pass. We initialize the predictor $f_{\boldsymbol{\theta}}$ from the SD2 U-Net and fully fine-tune it for single-step normal regression. The model predicts the clean normal latent conditioned on the RGB latent, semantic features, and the normal task embedding s_{n}:

$$\hat{\boldsymbol{z}}_{\text{n}} = f_{\boldsymbol{\theta}}(\boldsymbol{z}_{\text{rgb}}, T, \boldsymbol{c}_{\text{sem}}, s_{\text{n}}), \tag{6}$$

where T is a fixed timestep embedding. Similarly, the RGB reconstruction task predicts $\hat{\boldsymbol{z}}_{\text{rgb}} = f_{\boldsymbol{\theta}}(\boldsymbol{z}_{\text{rgb}}, T, \boldsymbol{c}_{\text{sem}}, s_{\text{rgb}})$. As illustrated in Fig. 2, the U-Net follows a standard encoder-decoder structure with skip connections. At each layer l, we flatten the spatial feature map into tokens $\boldsymbol{h}^{(l)} \in \mathbb{R}^{n \times d_l}$, where n is the number of spatial tokens and d_l is the feature channel dimension at layer l. We then apply cross-attention with $\boldsymbol{c}_{\text{sem}}$ as keys and values:

$$\begin{aligned} &\boldsymbol{Q} = \boldsymbol{h}^{(l)}\boldsymbol{W}_Q, \quad \boldsymbol{K} = \boldsymbol{c}_{\text{sem}}\boldsymbol{W}_K, \quad \boldsymbol{V} = \boldsymbol{c}_{\text{sem}}\boldsymbol{W}_V, \\ &\text{CrossAttn}(\boldsymbol{h}^{(l)}, \boldsymbol{c}_{\text{sem}}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}, \end{aligned} \tag{7}$$

where d_k is the attention head dimension. This design allows spatial features to query high-level semantics for geometry inference under ambiguous transparency cues.

4.3. Decoding & Wavelet Regularization

Decoding. The predicted latent is decoded to the pixel space in a single step:

$$\hat{\boldsymbol{N}} = \mathcal{D}_{\text{vae}}(\hat{\boldsymbol{z}}_{\text{n}}). \tag{8}$$

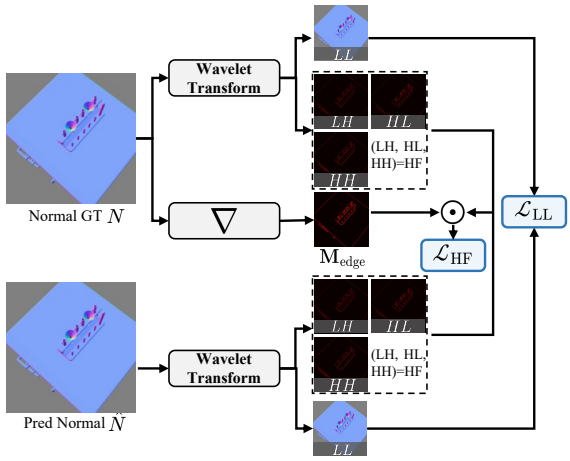


Figure 3. **Illustration of our Wavelet Edge-Aware Regularization.** We decompose both predicted and ground truth normals using the Haar wavelet transform into low-frequency (LL) and high-frequency (LH, HL, HH) sub-bands. An edge mask $\boldsymbol{M}_{\text{edge}}$ derived from GT normals enables spatially selective supervision: ① LL fidelity ensures correct overall shape; ② edge-aligned HF supervision preserves sharp boundaries only at true edges. (§ 4.3)

Training Losses. We use three losses: $\mathcal{L}_{\text{normal}}$, $\mathcal{L}_{\text{rgb}}$, and $\mathcal{L}_{\text{wavelet}}$. We define the latent reconstruction losses as:

$$\mathcal{L}_{\text{normal}} = \|\hat{\boldsymbol{z}}_{\text{n}} - \boldsymbol{z}_{\text{n}}\|_2^2, \quad \mathcal{L}_{\text{rgb}} = \|\hat{\boldsymbol{z}}_{\text{rgb}} - \boldsymbol{z}_{\text{rgb}}\|_2^2. \tag{9}$$

Wavelet Edge-Aware Regularization. Laboratory glassware exhibits a distinctive geometric prior: sharp normal discontinuities occur primarily at object boundaries and structural edges (*e.g.*, rims, bases, and liquid-glass interfaces), while interior regions exhibit smooth, continuous surfaces. Standard pixel-wise losses treat all regions uniformly, often over-smoothing edges to minimize global error. We address this through a wavelet-based regularization that provides edge-selective frequency supervision (Fig. 3).

Using the 2D Haar wavelet transform $\mathcal{W}$, we decompose both the predicted normal $\hat{\boldsymbol{N}}$ and ground truth $\boldsymbol{N}$:

$$\begin{aligned} \mathcal{W}(\hat{\boldsymbol{N}}) &= \{\hat{LL}, \hat{LH}, \hat{HL}, \hat{HH}\}, \\ \mathcal{W}(\boldsymbol{N}) &= \{LL, LH, HL, HH\}, \end{aligned} \tag{10}$$

where $LL \in \mathbb{R}^{3 \times \frac{H}{2} \times \frac{W}{2}}$ is the low-frequency approximation and $HF = [LH; HL; HH] \in \mathbb{R}^{9 \times \frac{H}{2} \times \frac{W}{2}}$ denotes the channel-wise concatenation of the three high-frequency sub-bands; predicted sub-bands use a hat, *e.g.*, $\hat{LL}$ and $\hat{HF}$. We define the edge mask $\boldsymbol{M}_{\text{edge}} = \frac{1}{2}\left(\|\nabla_x \boldsymbol{N}\|_2 + \|\nabla_y \boldsymbol{N}\|_2\right)$, normalized to $[0, 1]$ and downsampled to match the sub-band resolution, where ∇_x and ∇_y denote finite differences. The wavelet loss is then:

$$\begin{aligned} \mathcal{L}_{\text{LL}} &= \|\hat{LL} - LL\|_1, \\ \mathcal{L}_{\text{HF}} &= \|\boldsymbol{M}_{\text{edge}} \odot (\hat{HF} - HF)\|_1, \\ \mathcal{L}_{\text{wavelet}} &= \mathcal{L}_{\text{LL}} + \mathcal{L}_{\text{HF}}. \end{aligned} \tag{11}$$

Table 1. **Quantitative comparison on transparent object normal estimation**. We evaluate on ClearGrasp, our proposed TransNormal-Synthetic, and ClearPose datasets. Metrics: mean angular error (Mean↓, lower is better) and percentage of pixels within 11.25° and 30° thresholds (↑, higher is better). TransNormal achieves the best results across all three datasets. The best, second best, and third best results are highlighted. ⋆: diffusion-based; †: transformer-based. SA: SIGGRAPH Asia. (§ 5.4)

Method		Venue	ClearGrasp (Synthetic)			TransNormal-Synthetic			ClearPose (Real-World)			Avg.
			Mean↓	11.25° ↑	30° ↑	Mean↓	11.25° ↑	30° ↑	Mean↓	11.25° ↑	30° ↑	Rank
Omnidata	(Eftekhar et al.)	ICCV 21	36.9	15.1	49.1	11.3	80.9	89.3	48.3	10.8	33.8	12.3
Omnidata V2†	(Kar et al.)	CVPR 22	33.8	18.3	55.9	8.2	87.0	92.6	51.7	13.8	33.2	10.9
GeoWizard⋆	(Fu et al.)	ECCV 24	31.3	20.8	59.5	9.4	78.9	95.0	36.8	14.2	49.7	10.1
StableNormal⋆	(Ye et al.)	SA 24	32.0	17.5	65.3	7.6	86.8	96.3	37.1	14.1	57.5	8.9
Marigold⋆	(Ke et al.)	CVPR 24	27.6	31.0	65.3	6.2	90.4	96.3	33.0	25.5	57.5	6.3
DSINE	(Bae & Davison)	CVPR 24	25.7	26.4	68.6	13.2	70.3	90.7	40.2	15.9	46.3	9.6
Diff-E2E-FT⋆	(Martin Garcia et al.)	WACV 25	22.6	42.1	73.3	5.2	91.9	97.0	32.0	32.5	59.4	3.3
GenPercept⋆	(Xu et al.)	ICLR 25	25.8	30.3	70.9	6.9	87.6	97.0	31.6	31.2	63.0	4.2
Lotus-G⋆	(He et al.)	ICLR 25	21.7	39.7	75.4	8.2	82.3	96.7	31.8	28.8	60.4	5.2
Lotus-D⋆	(He et al.)	ICLR 25	21.9	37.0	75.7	9.0	80.9	97.1	31.3	23.2	59.5	5.3
MoGe-2†	(Wang et al.)	NeurIPS 25	26.6	17.0	64.2	6.2	90.1	96.8	36.2	14.3	48.3	7.8
Diception⋆	(Zhao et al.)	NeurIPS 25	29.5	25.8	65.3	7.1	88.3	97.3	31.0	33.8	63.5	5.0
TransNormal	**(Ours)**	-	16.4	51.7	85.0	4.1	93.5	98.2	26.3	35.9	69.8	1.0

Table 2. **Ablation on loss functions.** We evaluate different wavelet loss configurations: without wavelet, LL only, LL + interior HF, and our LL + edge HF. Our configuration achieves the best performance. (§ 5.4)

Loss Configuration	ClearPose		
	Mean↓	11.25° ↑	30° ↑
w/o $\mathcal{L}_{\text{wavelet}}$	29.1	30.0	64.1
LL only	29.4	29.4	64.1
LL + interior HF	27.6	33.5	67.1
LL + edge HF (Ours)	26.3	35.9	69.8

Table 3. **Ablation on semantic encoder.** We compare DINOv2, SigLIP2, SAM2, and DINOv3 as semantic guidance sources. DINOv3 (Ours) yields the best results. (§ 5.4)

Encoder	ClearPose		
	Mean↓	11.25° ↑	30° ↑
DINOv2	28.5	30.9	66.1
SigLIP2	27.2	34.5	67.8
SAM2	28.5	31.1	66.1
DINOv3 (Ours)	26.3	35.9	69.8

Table 4. **Ablation on fine-tuning strategies.** We compare: removing DINOv3 guidance entirely, applying LoRA to U-Net or DINOv3, versus our full U-Net fine-tuning with frozen DINOv3. Our strategy achieves the best performance. (§ 5.4)

Method	Fine-tuning		ClearPose		
	DINOv3	U-Net	Mean↓	11.25° ↑	30° ↑
w/o DINOv3	–	Full FT	27.7	33.4	67.2
U-Net LoRA	Frozen	LoRA	29.8	26.2	63.4
DINOv3 LoRA	LoRA	Full FT	27.5	34.7	67.5
Ours	Frozen	Full FT	26.3	35.9	69.8

The two terms target complementary geometric aspects: ① **Low-frequency fidelity**: supervises the LL sub-band to ensure correct overall shape and smooth curvature alignment with the ground truth. ② **Edge-selective high-frequency alignment**: enforces HF fidelity only at edges (weighted by M_{edge}), preserving sharp boundary reconstruction without introducing constraints on interior regions.

Total Loss. The final objective combines the normal/RGB losses with the wavelet regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{normal}} + \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{wv}}\mathcal{L}_{\text{wavelet}}. \quad (12)$$

5. Experiments

5.1. Implementation Details

We implement the proposed TransNormal by fine-tuning Stable Diffusion 2 (Rombach et al., 2022). During training, the VAE encoder and decoder are kept frozen, while the U-Net parameters and the linear projector are updated. The task embeddings s_{n} and s_{rgb} remain fixed. For the DINOv3 encoder, we use patch size $p = 16$. For optimization, we use the AdamW (Loshchilov & Hutter, 2019) optimizer with a learning rate of 3×10^{-5}. We apply random horizontal flipping for data augmentation during training. All models are trained on 8 NVIDIA A100 GPUs (80G) with a total batch size of 32 for 15,000 steps. During inference, we directly predict the normal map in a single inference step. For loss weights, we set $\lambda_{\text{rgb}} = 1.0$ and $\lambda_{\text{wv}} = 0.1$, with equal weights for the LL and edge high-frequency terms in the wavelet regularization.

5.2. Environment Setup

Training Data. This work aims to achieve strong performance using relatively limited supervised data. The normal estimation task is trained solely on a collection of synthetic data. During training, we sample from the following datasets with a ratio of **35:15:45:5**: ① *ClearGrasp* (Sajjan et al., 2020) (35%): a dataset for transparent objects containing 45,454 synthetic normal images; ② *TransNormal-Synthetic* (15%): a Blender-rendered dataset of laboratory scenes with transparent glassware introduced in this work, providing 3,555 training and 395 testing samples with pixel-accurate normals, depth, and segmentation masks (details in Appendix A); ③ *Hypersim* (Roberts et al., 2021) (45%): a photorealistic synthetic dataset of 461 indoor scenes, from which we utilize the official training split retaining 39,648 samples after filtering, resized to 576×768; ④ *Virtual KITTI* (Cabon et al., 2020) (5%): a synthetic street-scene dataset covering five urban scenes, from which we use four scenes comprising 33,580 samples, cropped to 352×1216.

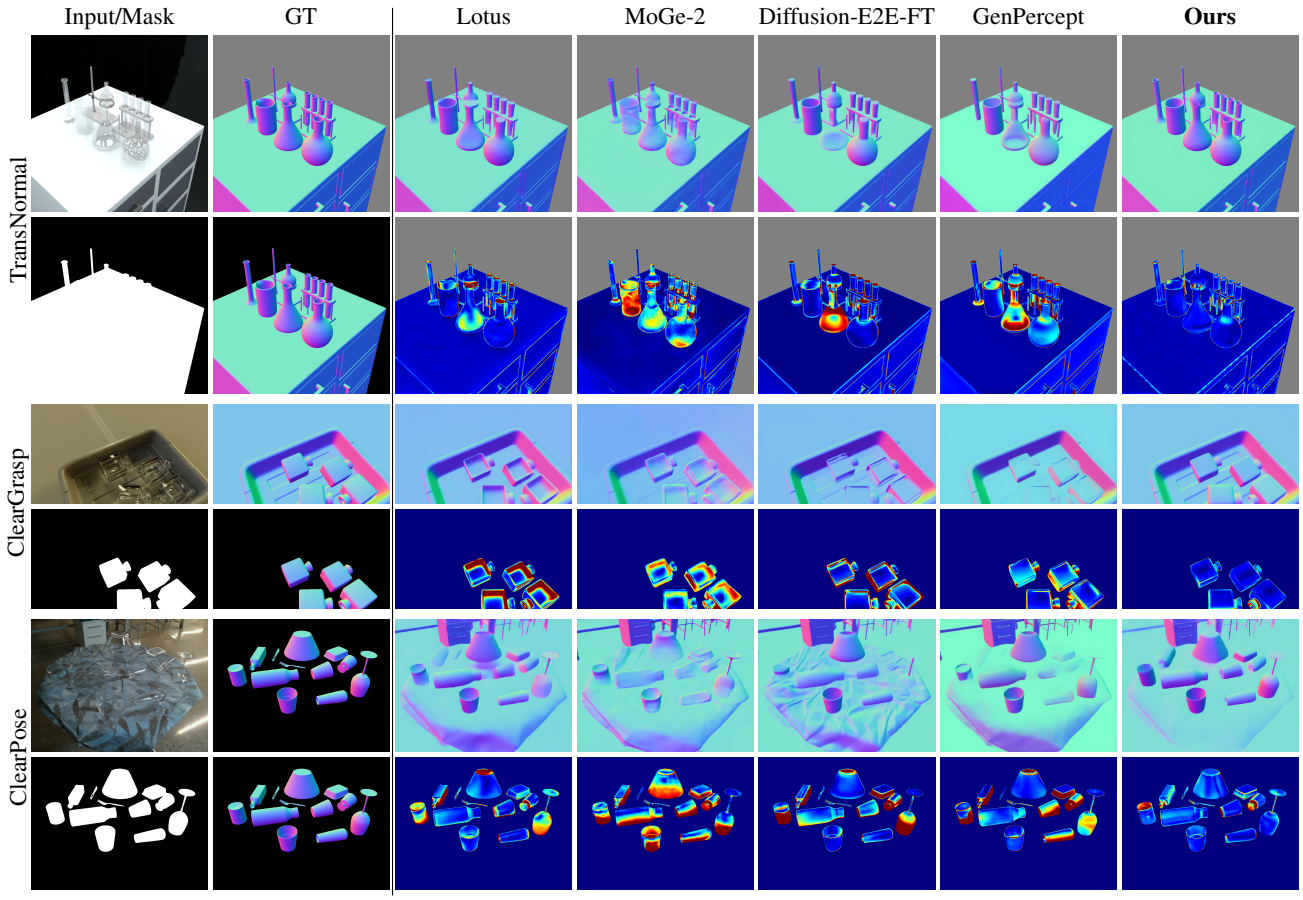


Figure 4. **Qualitative comparison on transparent object normal estimation.** We compare our method against state-of-the-art approaches across TransNormal-Synthetic, ClearGrasp, and ClearPose datasets. For each dataset, the top row shows predicted normals and the bottom row shows angular error maps (blue: low, red: high). Notably, even on ClearPose, an extremely challenging real-world dataset with diverse transparent objects under cluttered scenes, our method achieves superior zero-shot performance compared to other approaches. Existing methods produce blurry or incorrect normals on transparent regions due to refraction, while our method recovers sharp and accurate surface geometry. Please zoom in 🔍 for details. (§ 5.3)

Evaluation Data. We evaluate TransNormal on transparent object normal estimation using: the synthetic test split of *ClearGrasp* (Sajjan et al., 2020) (408 samples), the held-out test set of *TransNormal-Synthetic* (395 samples), and *ClearPose* (Chen et al., 2022) (120 samples). ClearPose is a challenging real-world dataset with diverse transparent objects under varying lighting conditions; we use it for zero-shot evaluation (not included in training) to assess generalization. For ClearPose, we use the subset with available meshes and recompute normals by reprojecting the ground-truth mesh, evaluating only within the transparent object mask. We apply this protocol to all compared methods.

Baselines. We compare TransNormal against representative normal estimation methods on the task of transparent object normal reconstruction. The baselines include models trained on opaque or general scenes (Omnidata (Eftekhar et al., 2021), Omnidata V2 (Kar et al., 2022), DSINE (Bae & Davison, 2024)) and diffusion-based dense prediction methods (GeoWizard (Fu et al., 2024), StableNormal (Ye et al., 2024), Marigold (Ke et al., 2024), Lotus (He et al., 2025), Diffusion-E2E-FT (Martin Garcia et al., 2025), GenPercept (Xu et al., 2025a), MoGe-2 (Wang et al., 2025b), Diception (Zhao et al., 2025)).

5.3. Qualitative Results

Comparison with Baselines. Fig. 4 presents qualitative comparisons between TransNormal and state-of-the-art methods across three transparent object benchmarks. For each dataset, the first row shows predicted normal maps, and the second row displays error maps within the transparent object mask (blue: low error, red: high error). Existing methods produce severely distorted normal predictions in transparent regions, as they are misled by refracted background textures. In contrast, TransNormal leverages DINOv3 semantic guidance to provide high-level shape understanding, enabling accurate geometry recovery even under challenging refractive conditions. Additional qualitative results are provided in Appendix C.1, and in-the-wild generalization

examples are shown in Appendix C.3.

5.4. Quantitative Results

Metrics. Following prior works (Bae & Davison, 2024; Ye et al., 2024; He et al., 2025), we measure the *mean angular error* (Mean↓) and the percentage of pixels within 11.25° and 30° thresholds (↑). The Avg. Rank is computed by ranking each method on every metric across all three datasets, then averaging all nine per-metric ranks.

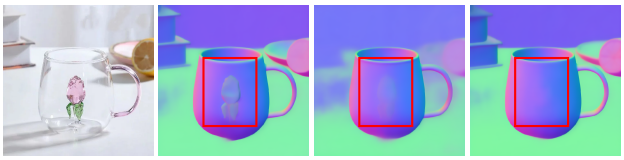
Results on ClearGrasp. On the synthetic ClearGrasp benchmark, TransNormal achieves a mean angular error of 16.4°, outperforming the previous best method Lotus-G (21.7°) by 24.4% relative improvement. Our method achieves 51.7% accuracy at the strict 11.25° threshold and 85.0% at 30°, indicating better fine-grained geometric accuracy. These results suggest that DINOv3 semantic guidance helps reduce ambiguities caused by refraction in transparent objects, where discriminative methods like DSINE (25.7°) and recent diffusion-based methods like Marigold (27.6°) struggle due to misleading local texture cues.

Results on TransNormal-Synthetic. Our proposed synthetic benchmark provides controlled evaluation of transparent object understanding. TransNormal achieves the best performance with 4.1° mean error and 93.5% accuracy at 11.25°, surpassing the strong baseline Diffusion-E2E-FT (5.2°, 91.9%). The consistent gains across metrics suggest that our semantic-guided architecture helps disentangle geometry from optical appearance, which is a key design principle of the TransNormal-Synthetic dataset.

Results on ClearPose. On the large-scale ClearPose dataset, TransNormal achieves the best results among the compared methods with 26.3° mean error and 69.8% accuracy at 30°, outperforming Diception (31.0°, 63.5%) and Lotus-D (31.3°, 59.5%). The 15.2% relative improvement in mean error suggests good generalization to diverse transparent object categories and poses. Traditional methods trained on opaque objects show large degradation (Omnidata V2: 51.7°), while our approach maintains best performance by leveraging semantic understanding to infer plausible geometry under challenging refractive conditions.

Ablation Studies. We conduct comprehensive ablation experiments on the ClearPose dataset to validate the effectiveness of our key design choices (Tab. 2, Tab. 3, Tab. 4, and Fig. 5).

① *Loss function design (Tab. 2, details in Appendix B.2).* Our wavelet-based loss design is important for transparent objects. Removing the wavelet regularization increases mean error from 26.3° to 29.1°, a 10.6% relative degradation. The spatially-selective frequency supervision is key: supervising only the LL sub-band lacks edge sharpness. The "LL + interior HF" configuration improves upon LL-only by suppressing spurious gradients in smooth regions, but still underperforms our full design that emphasizes edge-selective HF alignment.



(a) Input (b) w/o DINO (c) w/o Wavelet (d) Ours (Full)

Figure 5. **Qualitative ablation study on in-the-wild objects.** (a) In-the-wild input RGB image, a transparent cup with a flower inside. (b) Without DINOv3 semantic guidance, the model fails to recognize that the cup is transparent, incorrectly predicting the internal flower as surface geometry. (c) Without wavelet loss, the output exhibits discontinuous artifacts on smooth surfaces. (d) Our full model achieves both correct transparency understanding and smooth, continuous predictions. (§ 5.4)

② *Semantic encoder choice (Tab. 3, details in Appendix B.3).* We compare four vision encoders for semantic guidance. DINOv3 achieves the best results across all metrics, outperforming DINOv2 (Oquab et al., 2023) (28.5°), SigLIP2 (Tschannen et al., 2025) (27.2°), and Segment Anything Model 2 (SAM2) (Ravi et al., 2024) (28.5°). The superior performance of DINOv3 can be attributed to its stronger object-level semantic understanding, which is critical for inferring geometry from misleading optical cues.

③ *Fine-tuning strategies (Tab. 4, details in Appendix B.4).* Full fine-tuning (Full FT) of the U-Net with frozen DINOv3 encoder achieves the best performance (26.3° mean error). Removing DINOv3 guidance degrades performance to 27.7°, confirming the importance of semantic features. LoRA-based adaptation hurts performance for both U-Net and DINOv3, suggesting that bridging the domain gap requires sufficient model capacity and that fine-tuning the encoder on limited data risks overfitting.

6. Conclusion

We present TransNormal, a framework for transparent object normal estimation that elevates the task from low-level feature extraction to high-level scene understanding. By replacing the underutilized text conditioning in Stable Diffusion with dense DINOv3 visual semantics, we transform the cross-attention mechanism into a powerful semantic-injection channel that resolves geometric ambiguities caused by refraction and reflection. TransNormal achieves the best results among the compared methods across three transparent object benchmarks with an average rank of 1.0, using only ∼122K synthetic training samples (∼1.4% of MoGe-2's 8.9M). This supports the effectiveness of adapting generative priors with semantic guidance for specialized geometric tasks, and suggests a path toward more reliable embodied AI systems in laboratory automation.

Impact Statement

This paper advances geometric perception for transparent objects, with applications in laboratory automation that may accelerate scientific discovery. Our work does not involve human subjects or personal data, and the synthetic dataset is generated from 3D models without privacy concerns. We do not foresee negative societal consequences beyond those common to general advances in computer vision.

References

Agrawal, A., Roy, R., Duisterhof, B. P., Hekkadka, K. B., Chen, H., and Ichnowski, J. Clear-splatting: Learning residual gaussian splats for transparent object manipulation. In *RoboNerF: 1st Workshop on Neural Fields in Robotics (ICRA 2024)*, 2024.

Bae, G. and Davison, A. J. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9535–9545, 2024.

Bai, F., Zhang, H., Tao, T., Wu, Z., Wang, Y., and Xu, B. Picor: Multi-task deep reinforcement learning with policy correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6728–6736, Jun. 2023.

Bai, F., Li, Y., Chu, J., Chou, T., Zhu, R., Wen, Y., Yang, Y., and Chen, Y. Retrieval dexterity: Efficient object retrieval in clutters with dexterous hand. *arXiv preprint arXiv:2502.18423*, 2025a.

Bai, F., Zhao, R., Zhang, H., Cui, S., Zhang, S., bo xu, Han, L., Wen, Y., and Yang, Y. STAR: Efficient preference-based reinforcement learning via dual regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Cabon, Y., Murray, N., and Humenberger, M. Virtual kitti 2, 2020.

Cai, Y., Zhu, Y., Zhang, H., and Ren, B. Consistent depth prediction for transparent object reconstruction from rgb-d camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3459–3468, October 2023.

Chen, X., Zhang, H., Yu, Z., Opipari, A., and Jenkins, O. C. Clearpose: Large-scale transparent object dataset and benchmark. In *European Conference on Computer Vision (ECCV)*, pp. 381–396. Springer, 2022.

Dai, Q., Zhang, J., Li, Q., Wu, T., Dong, H., Liu, Z., Tan, P., and Wang, H. Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects. In *European Conference on Computer Vision (ECCV)*, pp. 374–391. Springer, 2022.

Deng, W., Campbell, D., Sun, C., Kanitkar, S., Shaffer, M. E., and Gould, S. Differentiable neural surface refinement for modeling transparent objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20268–20277, 2024.

Eftekhar, A., Sax, A., Malik, J., and Zamir, A. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10786–10796, 2021.

Eigen, D., Puhrsch, C., and Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 27, 2014.

Fang, H., Fang, H.-S., Xu, S., and Lu, C. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *IEEE Robotics and Automation Letters*, 7(3):7383–7390, 2022.

Fang, H.-S., Wang, C., Fang, H., Gou, M., Liu, J., Yan, H., Liu, W., Xie, Y., and Lu, C. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023.

Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., and Long, X. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision (ECCV)*, pp. 241–258. Springer, 2024.

Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S. A., Hu, V. T., and Ommer, B. Depthfm: Fast generative monocular depth estimation with flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3203–3211, 2025.

He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., and Chen, Y.-C. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

Hong, Y., Chen, J., Cheng, Y., Han, Y., Van Reeth, F., Claesen, L., and Liu, W. Cluedepth grasp: Leveraging positional clues of depth for completing depth of transparent objects. *Frontiers in Neurorobotics*, 16:1041702, 2022. doi: 10.3389/fnbot.2022.1041702.

Hu, W., Gao, X., Li, X., Zhao, S., Cun, X., Zhang, Y., Quan, L., and Shan, Y. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2005–2015, 2025.

Ichnowski, J., Avigal, Y., Kerr, J., and Goldberg, K. Dexnerf: Using a neural radiance field to grasp transparent objects, 2021.

James, S., Wohlhart, P., Kalakrishnan, M., Kalashnikov, D., Irpan, A., Ibarz, J., Levine, S., Hadsell, R., and Bousmalis, K. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12627–12637, 2019.

Jiang, X., Zhu, Z., Gao, T., and Guo, N. Ebfa-6d: End-to-end transparent object 6d pose estimation based on a boundary feature augmented mechanism. *Sensors*, 24 (23):7584, 2024. doi: 10.3390/s24237584.

Kar, O. F., Yeo, T., Atanov, A., and Zamir, A. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18963–18974, 2022.

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C., and Schindler, K. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024.

Kim, J., Jeon, M.-H., Jung, S., Yang, W., Jung, M., Shin, J., and Kim, A. Transpose: Large-scale multispectral dataset for transparent object. *The International Journal of Robotics Research*, 43(6):731–738, 2024.

Li, M., Pang, P., Fan, H., Huang, H., and Yang, Y. Tsgs: Improving gaussian splatting for transparent surface reconstruction via normal and de-lighting priors. In *ACM Multimedia*, 2025.

Li, Z., Long, X., Wang, Y., Cao, T., Wang, W., Luo, F., and Xiao, C. Neto: Neural reconstruction of transparent objects with self-occlusion aware refraction-tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18547–18557, 2023.

Lind, S. K., Triebel, R., and Krüger, V. Making the flow glow-robot perception under severe lighting conditions using normalizing flow gradients. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11195–11201. IEEE, 2024.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Martin Garcia, G., Abou Zeid, K., Schmidt, C., de Geus, D., Hermans, A., and Leibe, B. Fine-tuning image-conditional diffusion models is easier than you think. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 753–762. IEEE, 2025.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., and Susskind, J. M. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10912–10922, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, pp. 234–241. Springer, 2015.

Sajjan, S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., and Song, S. Clear grasp: 3d shape estimation of

transparent objects for manipulation. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 3634–3642. IEEE, 2020.

Scharstein, D. and Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

Shao, M., Xia, C., Yang, Z., Huang, J., and Wang, X. Transparent shape from a single view polarization image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9277–9286, 2023.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P. Dinov3, 2025.

Sulc, A., Sato, I., Goldluecke, B., and Treibitz, T. Towards monocular shape from refraction. In *British Machine Vision Conference (BMVC)*, 2021.

Sun, J.-M., Wu, T., Yan, L.-Q., and Gao, L. Nu-nerf: Neural reconstruction of nested transparent objects with uncontrolled capture environment. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 43(6), 2024. doi: 10.1145/3687757.

Sun, T., Zhang, G., Yang, W., Xue, J.-H., and Wang, G. Trosd: A new rgb-d dataset for transparent and reflective object segmentation in practice. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. doi: 10.1109/TCSVT.2023.3254665.

Tang, T., Liu, J., Zhang, J., Fu, H., Xu, W., and Lu, C. Rftrans: Leveraging refractive flow of transparent objects for surface normal estimation and manipulation. *IEEE Robotics and Automation Letters*, 9(4):3735–3742, 2024. doi: 10.1109/LRA.2024.3364837.

Tao, S., Xiang, F., Shukla, A., Qin, Y., Hinrichsen, X., Yuan, X., Bao, C., Lin, X., Liu, Y., Chan, T.-K., Gao, Y., Li, X., Mu, T., Xiao, N., Gurha, A., Rajesh, V. N., Choi, Y. W., Chen, Y.-R., Huang, Z., Calandra, R., Chen, R., Luo, S., and Su, H. Demonstrating gpu parallelized robot simulation and rendering for generalizable embodied ai with maniskill3. In *Proceedings of Robotics: Science and Systems*, 2025.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Tomasi, C. and Kanade, T. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., and Yang, J. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025a.

Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., and Yang, J. Moge-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Wen, Y., Lin, J., Zhu, Y., Han, J., Xu, H., Zhao, S., and Liang, X. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Advances in Neural Information Processing Systems*, 37: 41051–41075, 2024.

Woodham, R. J. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980.

Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., and Luo, P. Segmenting transparent objects in the wild. In *European Conference on Computer Vision (ECCV)*, 2020.

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., and Luo, P. Segmenting transparent object in the wild with transformer. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., and Shen, C. What matters when repurposing diffusion models for general dense perception tasks? In *International Conference on Learning Representations (ICLR)*, 2025a.

Xu, H., Wang, Y. R., Eppel, S., Aspuru-Guzik, A., Shkurti, F., and Garg, A. Seeing glass: Joint point-cloud and depth completion for transparent objects. In *Conference on Robot Learning (CoRL)*, pp. 827–838. PMLR, 2022.

Xu, S., Wei, S., Wei, Q., Geng, Z., Li, H., Shen, L., Sun, Q., Han, S., Ma, B., Li, B., Ye, C., Zheng, Y., Wang, N., Zhang, S., and Zhao, H. Diffusion knows transparency: Repurposing video diffusion for transparent object depth

and normal estimation. *arXiv preprint arXiv:2512.23705*, 2025b.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., and Zhao, H. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024b.

Ye, C., Qiu, L., Gu, X., Zuo, Q., Wu, Y., Dong, Z., Bo, L., Xiu, Y., and Han, X. Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (TOG)*, 2024.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., and Ma, Y. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Zhang, H., Opipari, A., Chen, X., Zhu, J., Yu, Z., and Jenkins, O. C. Transnet: Category-level transparent object pose estimation. In *European Conference on Computer Vision Workshops (ECCVW)*, 2022.

Zhang, L. and Agrawala, M. Transparent image layer diffusion using latent transparency. *ACM Trans. Graph.*, 43 (4), July 2024. ISSN 0730-0301. doi: 10.1145/3658150.

Zhao, C., Liu, M., Zheng, H., Zhu, M., Zhao, Z., Chen, H., He, T., and Shen, C. Diception: A generalist diffusion model for visual perceptual tasks. *arXiv preprint arXiv:2502.17157*, 2025.

Zhou, S., Wang, Z., and Ye, D. Novel view synthesis of transparent object from a single image. *Computer Graphics Forum*, 42(1):21–32, 2023. doi: 10.1111/cgf.14714.

Zhu, L., Mousavian, A., Xiang, Y., Mazhar, H., van Eenbergen, J., Desingh, K., and Fox, D. Rgb-d local implicit function for depth completion of transparent objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4649–4658, 2021.

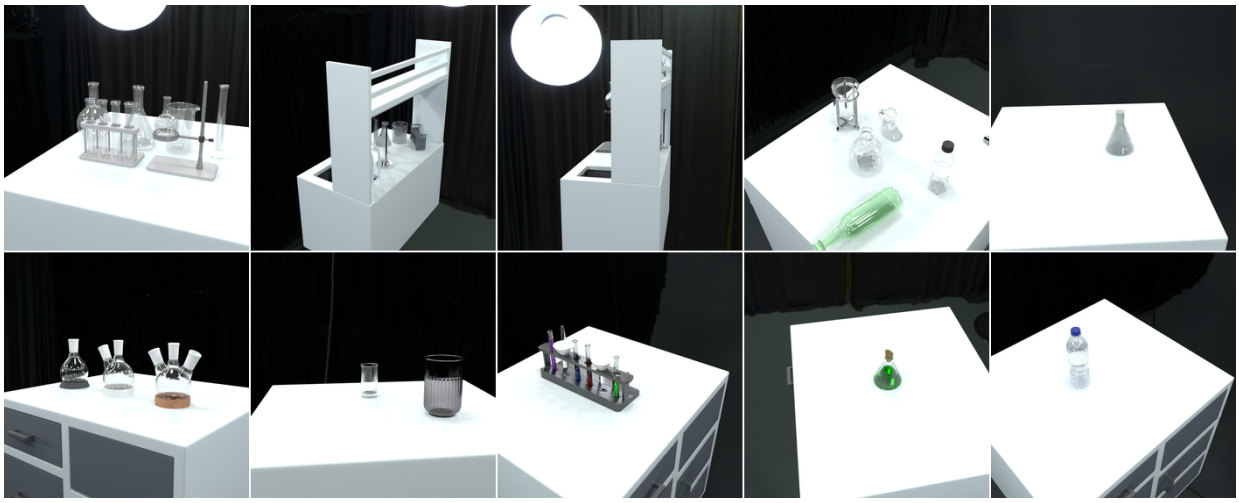


Figure 6. **Scene gallery of TransNormal-Synthetic.** Representative RGB renderings from different laboratory scenes, showcasing the diversity of transparent glassware configurations, lighting conditions, and background setups. (§ A)

A. The TransNormal-Synthetic Dataset

To address the scarcity of high-quality surface normal annotations for transparent objects, we introduce **TransNormal-Synthetic**, a curated synthetic dataset specifically designed for robust geometric perception. Leveraging the advanced physics-based rendering capabilities of Blender, we generate a diverse set of laboratory-style scenes containing ubiquitous transparent glassware such as beakers, test tubes, and pipettes. We will release the Blender scripts and `.blend` files (including various material presets), enabling users to construct custom datasets through simple scene composition.

A.1. Data Generation and Composition

TransNormal-Synthetic provides comprehensive multi-modal labels across 10 scenes, with 3,950 images in total. Each sample consists of the following components:

- **RGB Image Sequences**: To encourage invariance to optical appearance, each viewpoint includes three versions: (1) *RGB*, the standard rendering containing transparent objects; (2) *RGB with randomized material*, rendered by randomizing the transparent material parameters while keeping geometry fixed; and (3) *RGB background-only*, rendered by removing transparent objects to provide a clean reference.
- **Diverse Material Presets**: We provide multiple material options including translucent, fully transparent, and specular/glossy materials, enabling systematic evaluation under varying optical properties.
- **High-Precision Ground Truth**: We export pixel-accurate surface normal maps and 16-bit depth maps directly from the rendering engine. The depth maps are normalized following a 10m maximum distance protocol, consistent with laboratory-scale sensing.
- **Comprehensive Masks**: Each sample includes detailed segmentation masks, specifically identifying all objects (*foreground mask*) and specifically isolating transparent surfaces (*mask_transparent*).
- **Camera Parameters**: Full intrinsic matrices and 6D camera poses are provided to support potential downstream geometric reasoning tasks.

A.2. Material-Decoupled Design for Future Research

Beyond standard RGB-normal pairs, TransNormal-Synthetic provides a *material-decoupled* structure that enables future research on appearance-invariant geometry learning. By providing paired renderings that randomize transparent material parameters while keeping geometry fixed, this design can force a model to recognize that while the RGB appearance changes drastically with material variations, the underlying surface normal remains constant.

The inclusion of *RGB background-only* reference images further enables auxiliary tasks such as background inpainting, potentially leading to deeper understanding of light transport in refractive and scattering regions. While our current method uses only the standard RGB renderings, we release these additional modalities to support future exploration of material-invariant training strategies.

B. More Quantitative Results

B.1. Inference Efficiency

We benchmark TransNormal on a single NVIDIA A100 GPU, reporting average latency, FPS, and memory usage over 10 runs (Tab. 5). Mixed precision (BF16/FP16) yields $\sim$2.5$\times$ speedup over FP32, achieving 4.03 FPS. Peak memory is $\sim$11 GB, fitting within 16GB consumer GPUs.

Table 5. Inference efficiency of TransNormal (averaged over runs). Incremental memory is peak minus model load. (§ B.1)

Precision	Time (ms)	FPS	Peak Mem (MB)	Delta Mem (MB)	Model Load (MB)
BF16	247.98	4.03	11098.4	3642.2	7447.0
FP16	247.63	4.03	11098.0	3642.0	7447.0
FP32	615.43	1.63	10467.6	2200.1	8255.8

B.2. Loss Function Ablation Across Datasets

Tab. 6 evaluates our loss design across all benchmarks. We compare: (1) removing RGB reconstruction loss, (2) removing wavelet loss entirely, (3) supervising only LL sub-band, (4) LL + interior HF suppression, and (5) our full design with LL + edge-selective HF. Interior regions are defined as $(1 - \boldsymbol{M}_{\text{edge}})$, where $\boldsymbol{M}_{\text{edge}}$ is the normalized GT normal gradient.

Table 6. **Extended ablation on loss function design across three datasets.** We evaluate the contribution of each wavelet regularization component. The edge-selective high-frequency supervision (LL + edge HF) consistently outperforms alternatives. (§ B.2)

Loss Configuration	ClearGrasp						TransNormal-Synthetic						ClearPose					
	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑
w/o $\mathcal{L}_{\text{rgb}}$	16.7	17.9	32.2	50.2	77.2	85.1	4.7	76.2	88.9	93.2	97.0	98.1	26.7	10.1	19.1	32.5	59.4	69.1
w/o $\mathcal{L}_{\text{wavelet}}$	17.3	17.0	30.8	48.3	75.4	83.9	5.3	75.9	88.2	92.9	96.9	98.0	29.1	9.2	17.6	30.0	54.6	64.1
LL only	16.5	18.9	33.5	50.9	77.3	85.3	4.4	80.9	89.3	93.4	97.2	98.2	29.4	9.0	17.2	29.4	54.5	64.1
LL + interior HF	16.6	18.4	33.0	50.8	77.4	85.3	4.5	80.8	89.4	93.4	97.2	98.2	27.6	11.1	20.6	33.5	57.6	67.1
LL + edge HF (Ours)	16.4	19.7	34.4	51.7	77.2	85.0	4.1	84.1	90.3	93.5	97.1	98.2	26.3	11.0	21.6	35.9	61.0	69.8

B.3. Semantic Encoder Ablation Across Datasets

Tab. 8 compares four visual encoders, DINOv2, SigLIP2, SAM2, and DINOv3, across all three benchmarks, extending the analysis from Tab. 3. Tab. 7 lists the specific model variants and their specifications, including parameter counts, patch sizes, and feature dimensions.

Table 7. **Visual encoder specifications.** Model variants, parameter counts, patch sizes, and feature dimensions for the four encoders compared in the semantic encoder ablation (§ B.3).

Encoder	Model	Params	Patch Size	Feature Dim
DINOv2	dinov2-vitl14	304M	14	1024
SigLIP2	siglip2-large-patch16-384	304M	16	1024
SAM2	sam2-hiera-large	224M	16	256
DINOv3 (Ours)	dinov3-vith16plus	840M	16	1280

B.4. Fine-Tuning Strategy Ablation Across Datasets

Tab. 9 extends the fine-tuning strategy ablation from the main paper (Tab. 4) to all three benchmarks. We evaluate five configurations: (1) removing DINOv3 guidance entirely (using empty text prompt), (2) replacing DINOv3 with text prompt encoding (*e.g.*, "normal map"), (3) applying LoRA to the U-Net, (4) applying LoRA to DINOv3, and (5) our full model with frozen DINOv3 and fully fine-tuned U-Net.

Table 8. **Extended ablation on semantic encoder choice across three datasets.** We evaluate DINOv2, SigLIP2, SAM2, and DINOv3 (ours) as visual semantic guidance. DINOv3 consistently outperforms alternatives across both synthetic and real-world benchmarks. (§ B.3)

Encoder	ClearGrasp						TransNormal-Synthetic						ClearPose					
	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑
DINOv2	16.5	17.2	31.3	48.9	77.2	85.9	3.9	83.4	90.0	93.7	97.2	98.2	28.5	8.9	17.8	30.9	56.7	66.1
SigLIP2	16.7	18.0	31.8	49.2	76.9	85.3	4.7	74.0	90.3	93.8	97.3	98.3	27.2	11.0	21.3	34.5	58.7	67.8
SAM2	16.6	17.0	31.1	49.0	77.6	86.0	5.0	77.3	88.7	93.3	97.1	98.1	28.5	9.7	18.4	31.1	56.3	66.1
DINOv3 (Ours)	16.4	19.7	34.4	51.7	77.2	85.0	4.1	84.1	90.3	93.5	97.1	98.2	26.3	11.0	21.6	35.9	61.0	69.8

Table 9. **Extended ablation on fine-tuning strategies across three datasets.** We report mean angular error (Mean↓) and percentage of pixels within various angular thresholds (↑). Results demonstrate consistent trends across synthetic (ClearGrasp, TransNormal-Synthetic) and real-world (ClearPose) benchmarks. (§ B.4)

Method	Fine-tuning		ClearGrasp						TransNormal-Synthetic						ClearPose					
	DINOv3	U-Net	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑
w/o DINOv3	–	Full FT	16.6	16.8	31.2	49.2	77.2	85.7	4.5	78.8	90.3	93.8	97.3	98.2	27.7	10.8	20.1	33.4	58.1	67.2
Text Prompt	Text	Full FT	16.3	17.7	32.2	50.8	77.9	85.9	5.5	66.0	83.1	92.8	97.2	98.2	27.5	11.1	20.7	33.7	58.5	67.6
U-Net LoRA	Frozen	LoRA	16.3	17.4	32.0	50.2	78.5	86.4	5.7	61.4	83.7	92.3	96.8	97.8	29.8	6.6	14.0	26.2	53.0	63.4
DINOv3 LoRA	LoRA	Full FT	16.9	18.3	32.9	51.2	76.9	84.6	4.6	78.6	88.7	93.4	97.2	98.2	27.5	11.0	21.0	34.7	59.0	67.5
Full model (Ours)	Frozen	Full FT	16.4	19.7	34.4	51.7	77.2	85.0	4.1	84.1	90.3	93.5	97.1	98.2	26.3	11.0	21.6	35.9	61.0	69.8

B.5. Training Data Ratio Ablation

Our training combines ClearGrasp (CG) and TransNormal-Synthetic (TN)—both synthetic transparent object datasets with different object diversity and rendering characteristics. We ablate the CG:TN sampling ratio while keeping other data sources (Hypersim, Virtual-KITTI) fixed, evaluating five configurations from CG-dominant (45:5) to TN-dominant (20:30). Tab. 10 shows that our default 35:15 ratio achieves strong overall performance, particularly on ClearPose—a zero-shot evaluation benchmark—indicating better generalization to unseen real-world scenarios.

Table 10. **Ablation on training data sampling ratio.** We vary the balance between ClearGrasp (CG) and TransNormal-Synthetic (TN)—both synthetic transparent object datasets—while keeping other data sources fixed. Results show that our default ratio (35:15) achieves strong performance, while the sensitivity to exact ratios is relatively low. (§ B.5)

CG:TN	ClearGrasp						TransNormal-Synthetic						ClearPose					
	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑	Mean↓	5° ↑	7.5° ↑	11.25° ↑	22.5° ↑	30° ↑
40:10	16.4	18.1	32.3	50.5	77.8	85.6	4.4	81.6	89.1	93.2	97.0	98.1	26.8	10.8	20.9	34.5	59.7	68.6
25:25	16.5	18.4	33.0	51.4	77.9	85.4	4.2	82.2	89.8	93.7	97.3	98.3	26.7	11.0	20.6	33.9	59.2	68.5
20:30	16.9	18.2	33.3	51.3	77.1	84.6	5.0	73.0	88.4	93.4	97.2	98.2	28.5	10.5	19.5	32.5	57.0	66.2
45:5	16.9	19.1	33.2	49.9	76.0	84.4	4.5	80.1	90.1	93.6	97.2	98.1	27.1	11.1	20.7	34.0	58.7	67.9
35:15 (Ours)	16.4	19.7	34.4	51.7	77.2	85.0	4.1	84.1	90.3	93.5	97.1	98.2	26.3	11.0	21.6	35.9	61.0	69.8

C. More Qualitative Results

C.1. Extended Baseline Comparisons

Analysis. Fig. 10 and Fig. 11 present additional comparisons with 9 baseline methods on TransNormal-Synthetic and ClearGrasp. Across all examples, we observe consistent trends: ① baseline methods tend to over-smooth edges due to the lack of semantic guidance for distinguishing object boundaries from refracted backgrounds; ② methods without wavelet regularization produce blurred predictions on interior surfaces; ③ TransNormal maintains sharp edge reconstruction while preserving smooth interior surfaces, validating our design choices.

C.2. DINOv3 Semantic Feature Visualization

A core claim of TransNormal is that DINOv3 semantic features help resolve the *appearance-geometry decoupling* problem in transparent objects: refraction and transmission cause local RGB appearance to be dominated by background imagery rather than the object's intrinsic geometry. To validate this, we visualize the dense patch tokens extracted from DINOv3's final layer using Principal Component Analysis (PCA). The first three principal components are mapped to RGB channels, producing a colorized representation where similar colors indicate semantically similar regions. As shown in Fig. 12(b), DINOv3 features cluster by object structure—the eyewear forms coherent semantic groups distinct from the background—despite the transparent material causing the background to be visible through the lenses. This object-level semantic understanding

enables our method to correctly infer surface geometry.

C.3. Additional In-the-Wild Results

To evaluate whether TransNormal generalizes beyond laboratory glassware, we conduct zero-shot inference on in-the-wild transparent objects. Since ground truth is unavailable for these images, we perform qualitative comparison against 6 baselines: Lotus-D, DSINE, Diception, GeoWizard, Marigold, and MoGe-2 (Fig. 13).

D. Limitations and Future Work

While TransNormal significantly advances transparent object normal estimation, several directions warrant further exploration:

Multi-view and Temporal Consistency. Our current framework focuses on single-view estimation. Incorporating multi-view consistency constraints or temporal coherence for video sequences could further improve robustness and enable applications in dynamic manipulation scenarios.

Generalization to Other Dense Prediction Tasks. The semantic-guided architecture demonstrates strong performance on normal estimation. Exploring its generalization to other dense prediction tasks such as depth estimation, optical flow, or material property prediction represents a promising avenue for future research.

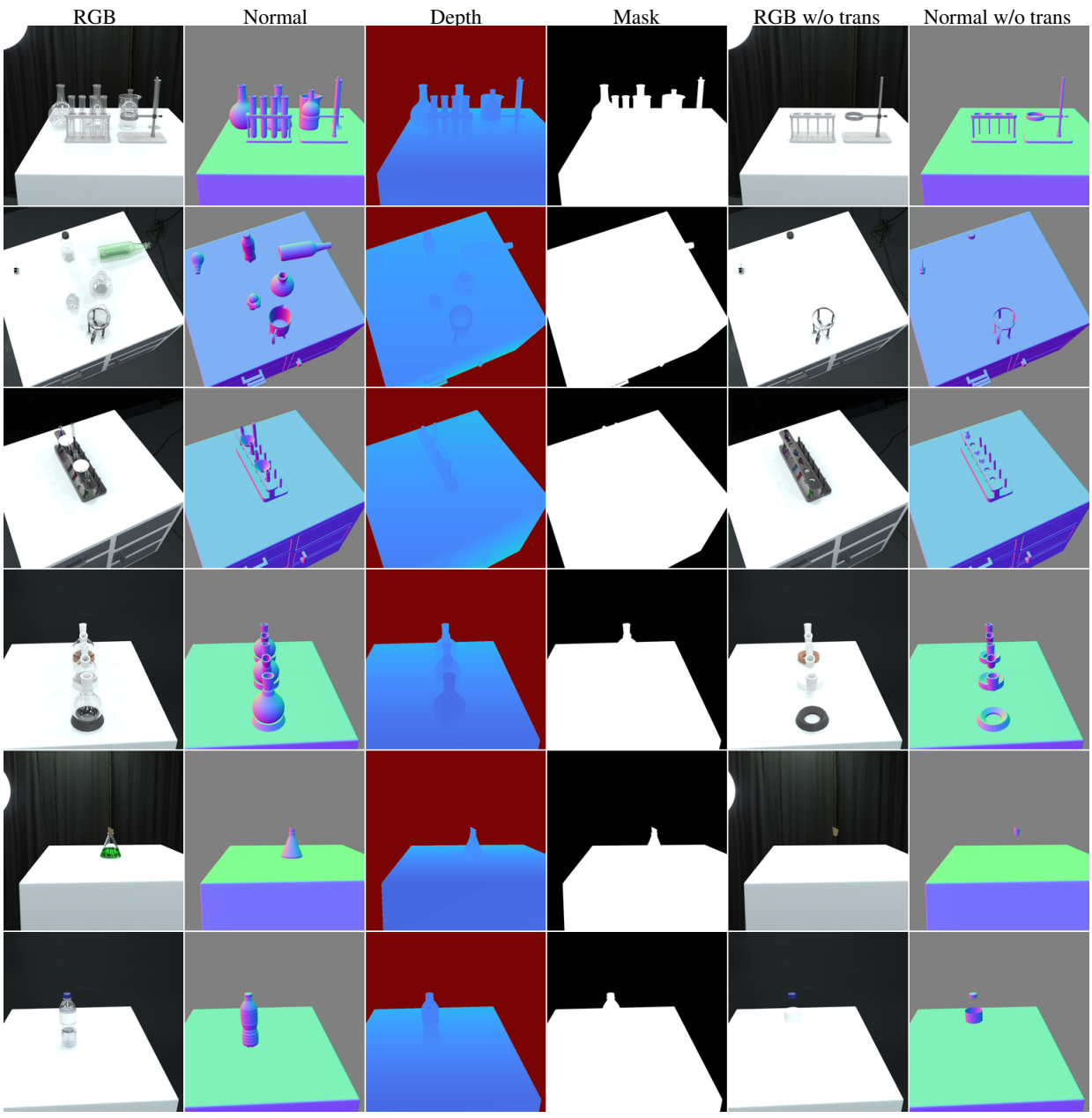


Figure 7. **Multi-modal annotations in TransNormal-Synthetic.** Each row shows a different scene with six annotation types. The material-decoupled design (with/without transparent objects) enables the model to learn geometry invariant to optical appearance. (§ A.1)

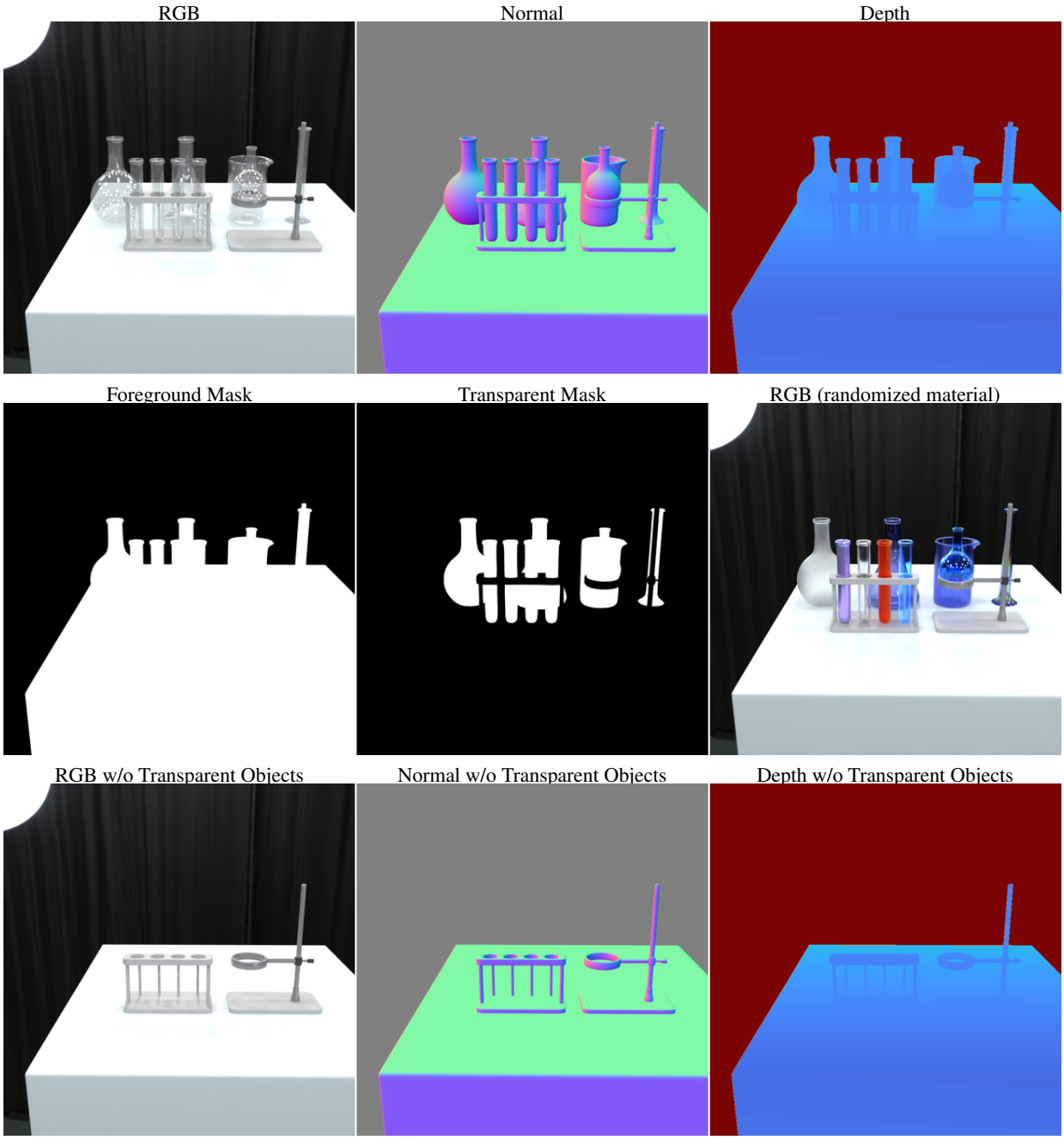


Figure 8. **Annotation detail visualization.** (Row 1) Standard rendering with transparent objects; (Row 2) Foreground mask, transparent mask, and RGB with randomized transparent material; (Row 3) Reference rendering without transparent objects. This triplet structure enables geometry-appearance disentanglement. (§ A.1)

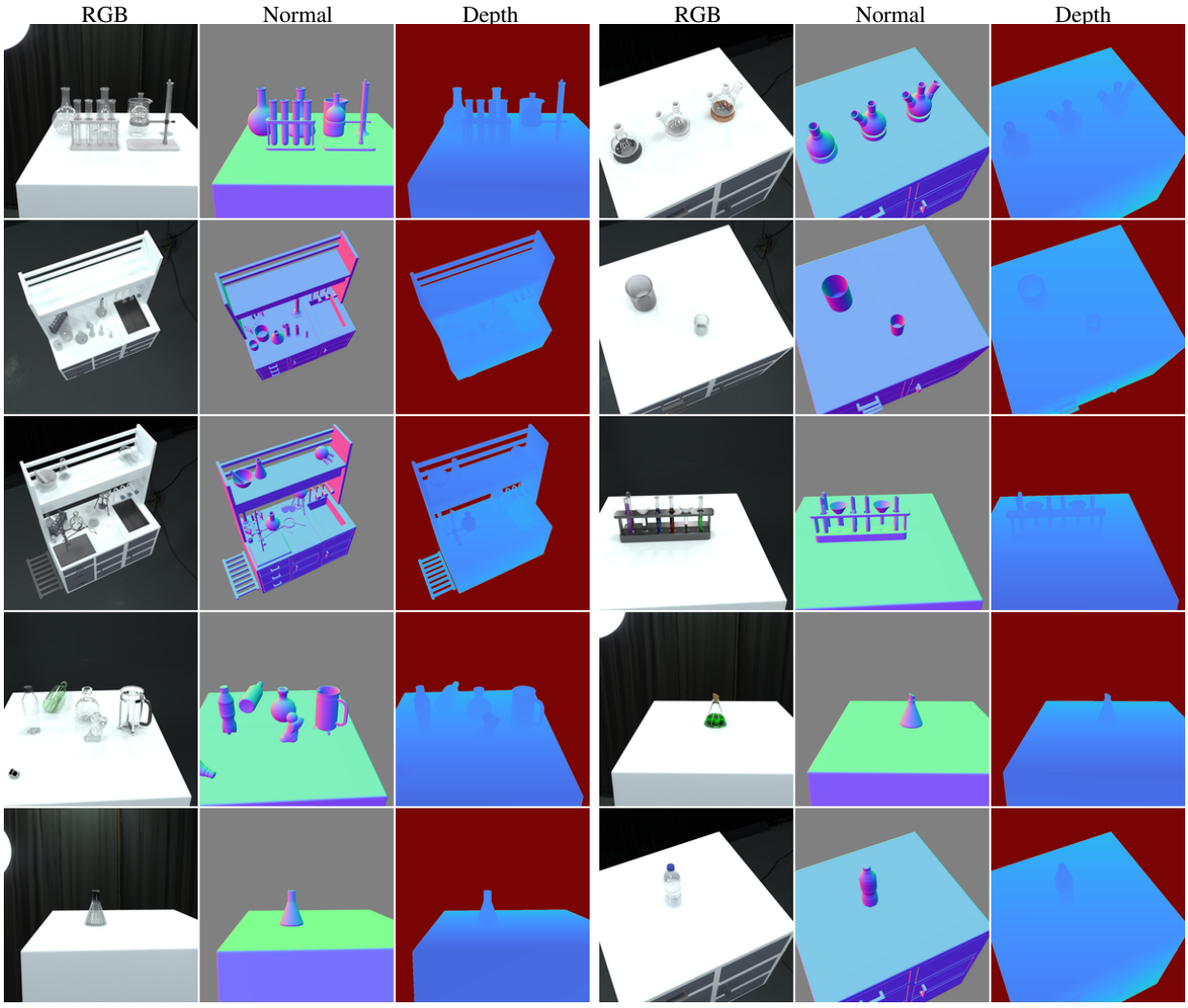


Figure 9. **Comprehensive scene coverage in TransNormal-Synthetic.** RGB images, surface normals, and depth maps across 10 representative scenes, demonstrating the dataset's coverage of diverse transparent object arrangements, viewpoints, and lighting conditions. (§ A)

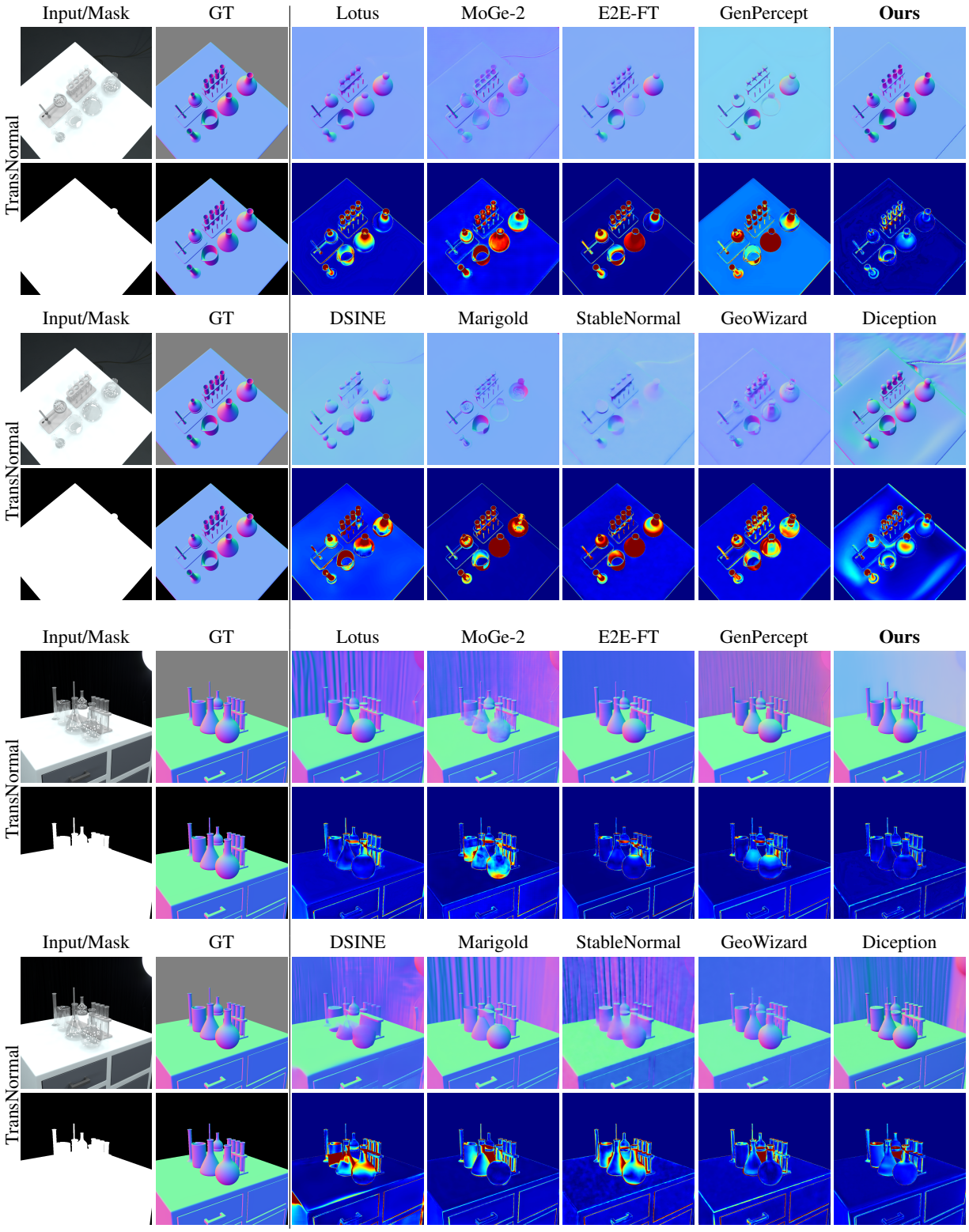


Figure 10. **Extended qualitative comparison with baseline methods.** We compare against 9 baselines. Top rows show predicted normals; bottom rows show angular error maps (blue: low, red: high). Our method consistently produces sharper edges and lower error on transparent regions. Please zoom in 🔍 for details. (§ C.1)

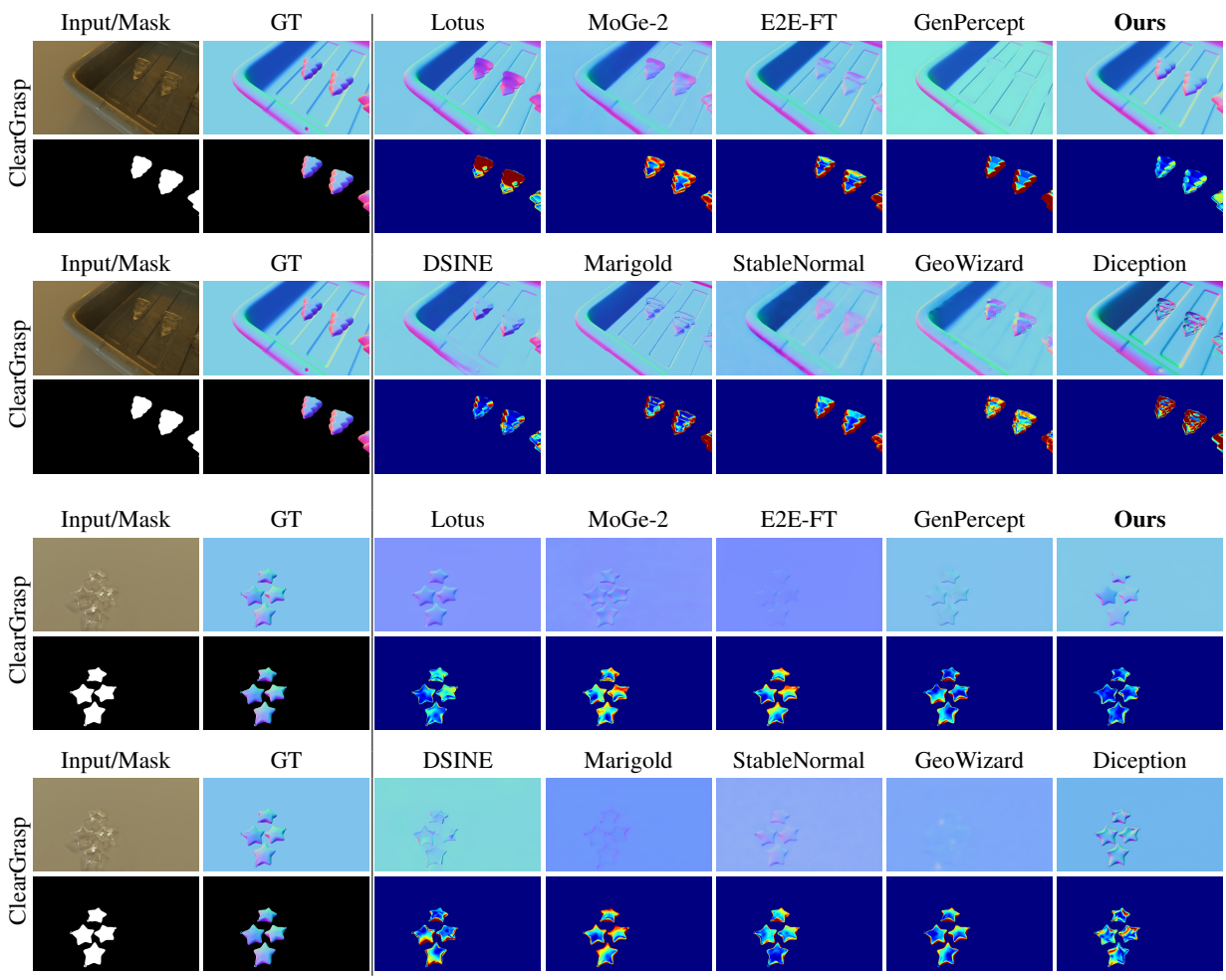
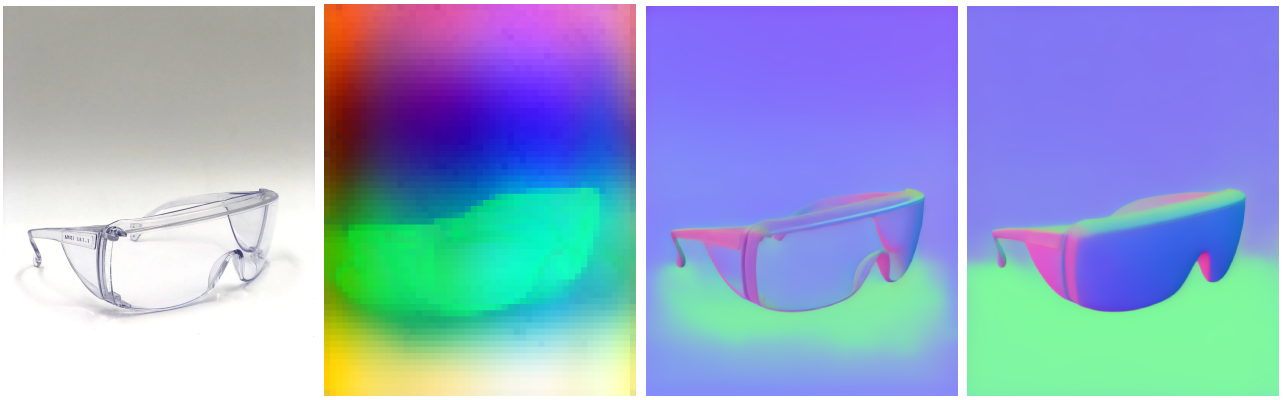


Figure 11. **Extended qualitative comparison with baseline methods.** We compare against 9 baselines. Top rows show predicted normals; bottom rows show angular error maps (blue: low, red: high). Our method produces sharper edges and lower error on transparent regions. Please zoom in 🔍 for details. (§ C.1)



(a) Input (b) DINOv3 (PCA) (c) Lotus-D (d) Ours

Figure 12. **DINOv3 semantic features capture object-level geometry priors.** (a) Input RGB image of transparent safety glasses exhibiting refraction and transmission; (b) DINOv3 patch tokens visualized via PCA—semantic features cluster by object structure rather than local texture, encoding canonical shape priors that distinguish the eyewear from refracted background textures and transmission artifacts; (c) Lotus-D struggles with transparent surfaces, producing noisy predictions affected by shadows and transmitted background imagery; (d) Our method leverages DINOv3 semantics to correctly recover smooth surface geometry. (§ C.2)

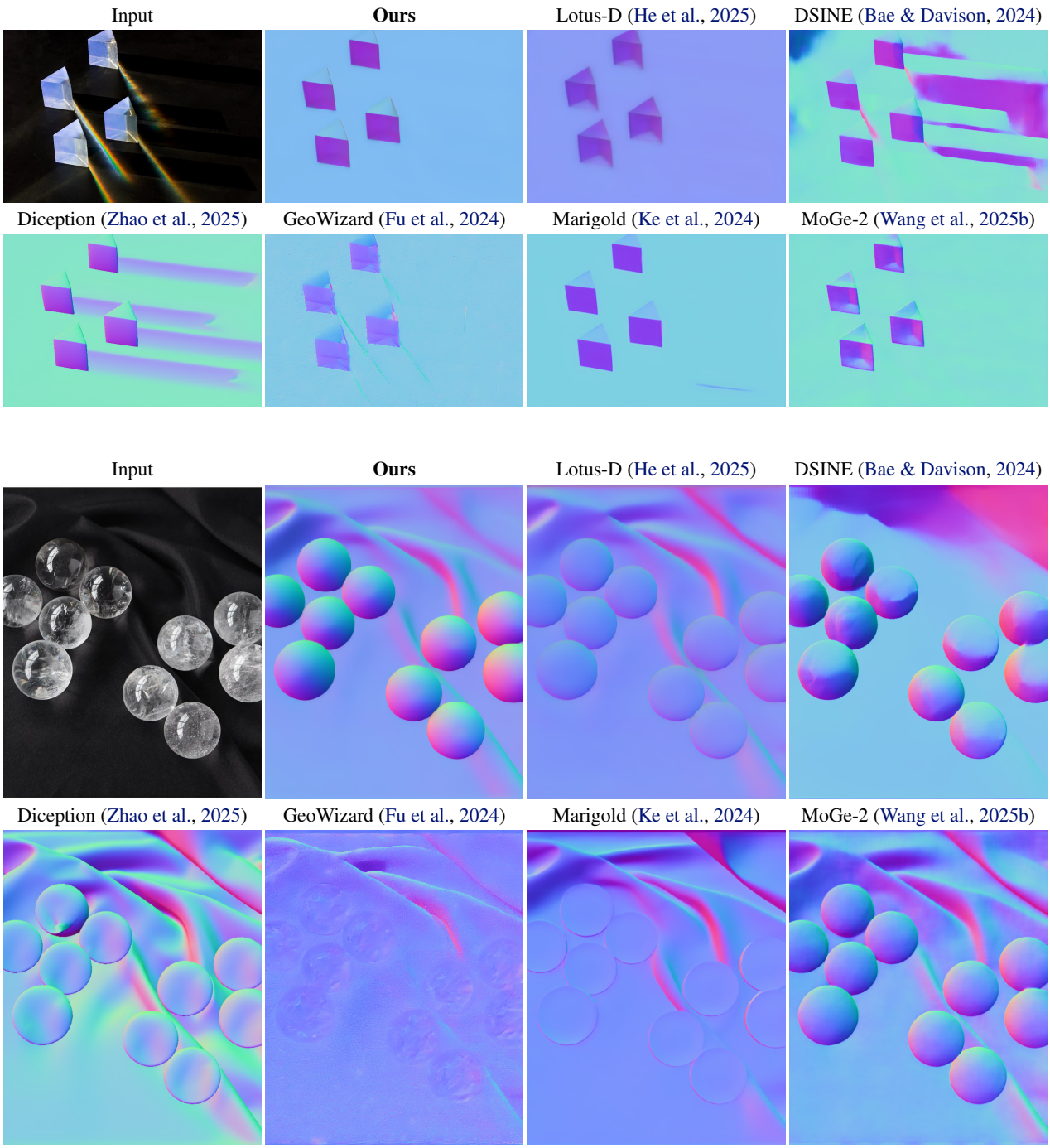


Figure 13. **Additional qualitative results on in-the-wild images.** We evaluate TransNormal on in-the-wild transparent objects and compare with 6 baselines. TransNormal produces more coherent surface normals on transparent regions, while baselines tend to be misled by refracted background textures or produce over-smoothed predictions. (§ C.3)