2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Diffusion Knows Transparency: Repurposing Video Diffusion for Transparent Object Depth and Normal Estimation

Shaocong Xu[1], Songlin Wei[2], Qizhe Wei[1], Zheng Geng[1], Hong Li[1,4], Licheng Shen[3], Qianpu Sun[3], Shu Han[5]
Bin Ma[3], Bohan Li[6,7], Chongjie Ye[8], Yuhang Zheng[9], Nan Wang[1], Saining Zhang[1], and Hao Zhao[1,3]

Abstract— **Transparent objects remain notoriously hard for perception systems: refraction, reflection and transmission break the assumptions behind stereo, ToF and purely discriminative monocular depth, causing holes and temporally unstable estimates. Our key observation is that modern video diffusion models already synthesize convincing transparent phenomena, suggesting they have internalized the optical rules. We build TransPhy3D, a synthetic video corpus of transparent/reflective scenes: 11k sequences (1.32M frames) rendered with Blender/Cycles. Scenes are assembled from a curated bank of category-rich static assets and shape-rich procedural assets paired with glass/plastic/metal materials. We render RGB + depth + normals with physically based ray tracing and OptiX denoising. Starting from a large video diffusion model, we learn a video-to-video translator for depth (and normals) via lightweight LoRA adapters. During training we concatenate RGB and (noisy) depth latents in the DiT backbone and co-train on TransPhy3D and existing frame-wise synthetic datasets, yielding temporally consistent predictions for arbitrary-length input videos. The resulting model, DKT, achieves zero-shot SOTA on real and synthetic video benchmarks involving transparency: ClearPose, DREDS (CatKnown/CatNovel), and TransPhy3D-Test. It improves accuracy and temporal consistency over strong image/video baselines (e.g., Depth-Anything-v2, DepthCrafter), and a normal variant (DKT-Normal) sets the best video normal estimation results on ClearPose. A compact 1.3B version runs at 0.17 s/frame (832×480). Integrated into a grasping stack, DKT's depth boosts success rates across translucent, reflective and diffuse surfaces, outperforming prior estimators. Together, these results support a broader claim: "Diffusion knows transparency." Generative video priors can be repurposed, efficiently and label-free, into robust, temporally coherent perception for challenging real-world manipulation. Code and models are available at `https://daniellli.github.io/projects/DKT/`.**

I. INTRODUCTION

Accurate depth estimation of transparent and reflective objects is fundamental to advancing 3D reconstruction [1] and robotic manipulation [2]. Nevertheless, the intrinsic physical ambiguities of these objects impose substantial limitations on depth-sensing cameras that rely on time-of-flight measurements or stereo correspondence [3], [4]. In particular, transparent objects often produce missing regions in depth maps, which in turn lead to degraded performance in downstream tasks.

Recent data-driven approaches have sought to address this challenge by constructing datasets [5] that encompass diverse lighting conditions and material properties, thereby approximating the visual characteristics of transparent and specular objects, and subsequently training models for depth prediction [4], [6]. However, such datasets remain constrained in diversity, and the resulting methods frequently exhibit suboptimal performance in real-world scenarios. We hypothesize that these methods tend to overfit to the limited datasets on which they are trained. To address the generalization challenge, recent works [4], [7] have increasingly leveraged pre-trained vision encoders, such as DINO [8], or harnessed text-to-image foundation models like Stable Diffusion [9] to train depth estimation networks. While these approaches have achieved notable improvements in single-frame depth accuracy, they continue to suffer from a lack of temporal consistency across frame sequences [10]. This limitation is particularly detrimental to downstream tasks that rely on stable 3D perception to support consistent action policies, such as robotic manipulation [11], [12]. These tasks are often carried out in dynamic and unstructured environments, where robust perception and temporally coherent decision-making are indispensable.

With recent advances in Video Diffusion Model (VDM) [13], [14], we observe their remarkable capacity to synthesize physically plausible videos of interactions with transparent objects, as illustrated in the first column of Fig. 1. Our central insight is that these models appear to have implicitly internalized the physical principles of light transport—such as refraction and reflection through transparent or translucent materials. To leverage this knowledge for video depth estimation of transparent-object, we make contributions from two perspectives: data and learning.

Data. We collect a 3D asset collection consisting of diverse categories and shapes of transparent and highly reflective items. Subsequently, we introduce a rendering pipeline that automatically generates physically plausible scenes using these assets and renders video data with varied light sources and camera trajectories, leading to the first synthetic video dataset, termed *TransPhy3D*, which focuses on transparent-objects, complementing existing image counterparts [4], [15], [16] that primarily study single-frame depth estimation problem. **Learning.** We propose a paradigm shift to video depth estimation: reframing it from a discriminative estimation task to a video-to-video translation problem. We

[1]Beijing Academy of Artificial Intelligence, scxu@baai.ac.cn
[2]University of Southern California.
[3]Tsinghua University, zhaohao@air.tsinghua.edu.cn.
[4]Beihang University.
[5]Wuhan University.
[6]Shanghai Jiao Tong University.
[7]European Institute of Innovation and Technology Ningbo.
[8]FNii, The Chinese University of Hong Kong, Shenzhen.
[9]National University of Singapore.

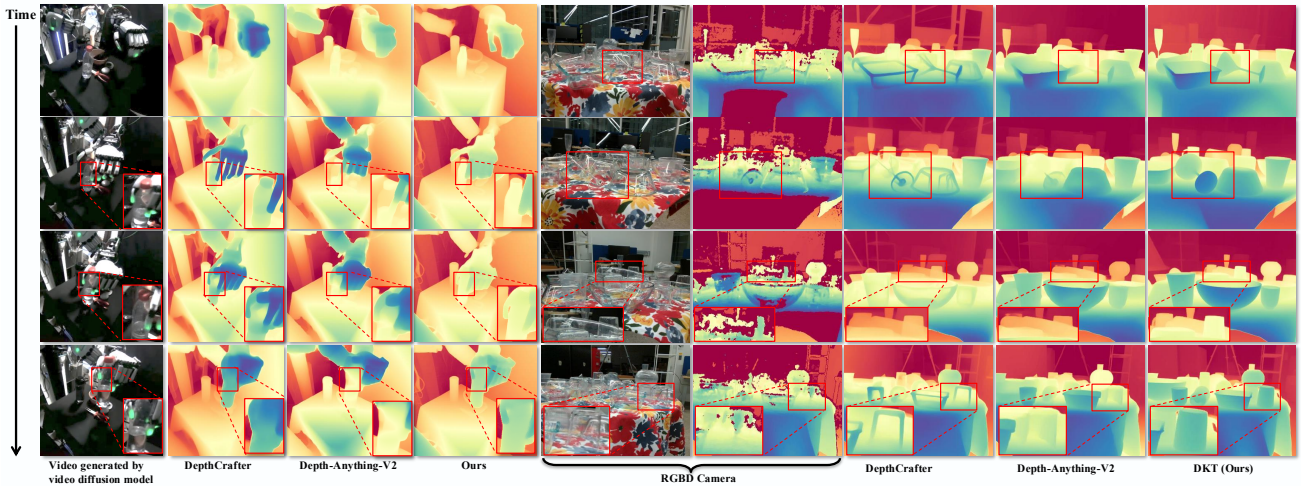

Fig. 1. We present DKT, a foundational model for fine-grained, temporally consistent depth estimation of in-the-wild videos featuring transparent objects of arbitrary lengths.

achieve this by repurposing VDM [14] using a LoRA training strategy. To fully leverage existing frame-wise datasets, we introduce a co-training strategy that enables joint training on a mixture of frame-wise and video data. Finally, we introduce a foundation model designed primarily for video depth estimation of transparent-object, termed **DKT**.

We validate the effectiveness of DKT through comprehensive experiments, demonstrating that it achieves SOTA performance under zero-shot setting on both synthetic and real-world benchmarks.

In summary, our main contributions are:

- We introduce TransPhy3D, the first synthetic transparent-object video dataset, comprising 11,000 videos and 1.32 million frames of data, to enable effective fine-tuning of VDM.
- We introduce the first foundation model for transparent-object video depth estimation by repurposing VDM through LoRA finetuning and we design a co-training strategy for training on mixture data of available synthetic image datasets and TransPhy3D.
- We conduct comprehensive benchmarking of existing SOTA methods on several open datasets and demonstrate the superiority of DKT in both depth estimation accuracy and real-world robotic experiments.

II. RELATED WORKS

A. From Discriminative to Generative Depth Estimation

Depth estimation has long been dominated by discriminative approaches [17], [18], [19], [4], [10], [17], [20], [21], [22], [23], [24], [25], [24], [26]. Early methods [17], [16] relied on synthetic data and hand-crafted geometric cues (e.g., surface normals, occlusion boundaries) to overcome the ambiguous visual appearance of transparency. Despite their innovation, these models suffered from significant domain gaps and limited generalization. Subsequent work incorporated stereo cues [26], [19], probabilistic volumetric representations [27], and metric learning techniques [25], [24], yet still operated within a discriminative framework, attempting to directly map pixels to depth. A turning point emerged with the adoption of generative models—especially diffusion models—which reframe depth estimation not as regression, but as a conditional generative process. Methods [4], [27] leverage diffusion to iteratively refine depth predictions, incorporating physical constraints such as stereo consistency and temporal smoothness. These approaches implicitly learn optical priors, enabling more robust inference on challenging materials. More recent video-depth techniques [28], [29], [10] further demonstrate that generative architectures inherently capture scene dynamics and material properties. Our work builds upon this generative turn, but goes further: we show that large-scale video diffusion models, pre-trained on internet-scale video data, already internalize a rich prior of transparent phenomena. By fine-tuning such a model *entirely on synthetic data*, we achieve SOTA zero-shot depth estimation without any real-world labels.

B. The Rise of Generative Data and Physics-Aware Synthesis

Parallel to advances in model architecture, the synthesis of training data has also undergone a generative revolution [16], [4], [15], [30], [31], [11], [32], [33], [34], [35], [36], [37], [38], [39], [40]. Early synthetic datasets for transparency, such as those produced by [16] and [15], relied on physically-based rendering (PBR) and careful domain randomization. While effective, these methods required significant expertise and computational resources to simulate realistic sensor noise and material variations. The advent of generative video models has introduced a new paradigm: models such as TI2V-Zero [30] can produce photorealistic, temporally coherent videos of transparent objects *without* fine-tuning, implying that generative priors encapsulate complex optical laws. Subsequent tools [41], [33] further enable fine-grained control over transparency effects in generated content. These capabilities suggest that generative models have learned not only appearance, but underlying physical rules. In robotics, this shift has enabled policies and perception systems that leverage generative world models, as seen in [11], [32]. Unlike traditional simulation-based data generation, generative models offer scalability and diversity, reducing reliance on hand-engineered graphics pipelines. Our approach embraces

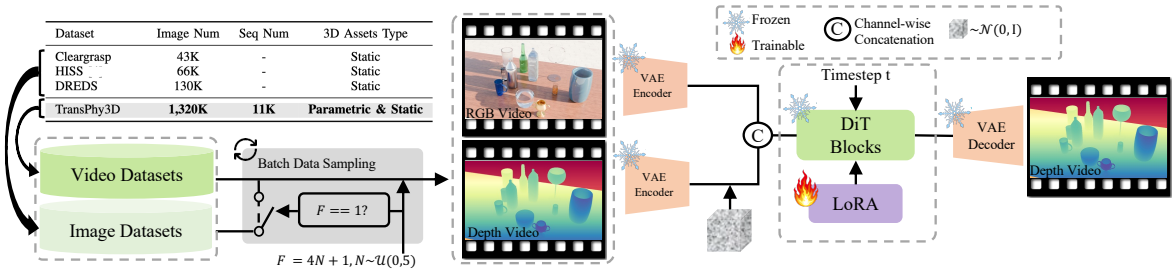

Fig. 2. **Overaview of DKT.** DKT starts with a pretrained video diffusion model [14] and is finetuned for video depth estimation by concatenating an extra RGB latent with the input latent using LoRA training strategy.

this idea: we use a generative video diffusion model as both a data synthesizer and a perception backbone. By fine-tuning it on purely synthetic transparent sequences, we exploit its inherent physical understanding—achieving superior generalization without real-world supervision.

III. METHOD

Despite the rapid progress in video depth and normal estimation [17], [42], [10], [43], existing methods remain limited when handling transparent and reflective objects. The core difficulty lies in the absence of reliable supervision: for real-world data, the SOTA pipeline [5]—RGB-D capture followed by CAD model recovery—fails to account for background information, while for synthetic data, no video dataset with transparent objects is available. This scarcity of ground truth highlights the need for strong generative priors. However, adapting generative prior models such as VDMs to this specific domain introduces a critical challenge—catastrophic forgetting of their original priors.

To tackle these challenges, we propose DKT, a framework that couples large-scale video data curation of transparent and reflective objects with LoRA-based adaptation of VDMs.

To mitigate the lack of supervision, we first construct a 3D asset bank with diverse categories and shapes, and introduce a rendering pipeline that generates physically plausible video scenes under varied lighting and camera trajectories. To further reduce rendering costs and enhance training efficiency, we incorporate existing synthetic image datasets and devise a heuristic sampling strategy that enables joint training on both image and video data within a unified pipeline. To address the scenario of catastrophic forgetting, we employ the LoRA strategy [44] to efficiently adapt VDM for transparency perception, achieving a seamless fusion of its transparent priors with the essential knowledge required for new tasks.

A. TransPhy3D

To address the gap in video datasets featuring transparent and reflective objects, we construct the first synthetic video dataset of transparent and reflective objects. This dataset is characterized by diversity in object shapes and categories, varied camera trajectories, and high-quality annotations.

Parametric & Static 3D Assets. As illustrated in Fig. 3, our asset repository integrates two complementary sources: *Category-Rich Static 3D Assets* and *Shape-Rich Parametric 3D Assets*, ensuring rich categories and shapes. For the former, we collected 5,574 assets from BlenderKit[1]. Each

[1]https://www.blenderkit.com/

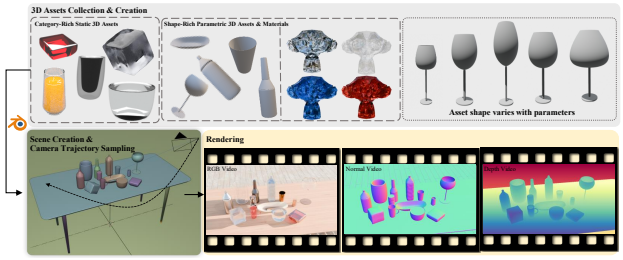

Fig. 3. **Rendering Pipeline.** Scenarios are constructed using static and parametric 3D assets. RGB, depth, and normal videos are rendered by sampling a circular trajectory within the scene.

asset is assigned an aesthetic score by rendering an image and passing it through Qwen2.5-VL-7B [45] to identify objects with transparent or highly reflective properties. Consequently, this process results in a final collection with 574 high-quality assets that is rich in categories, featuring transparent and reflective assets. For the latter, following [46], we develop a procedural pipeline to generate parametric assets. As shown in the bottom of Fig.3, varying parameters of the same asset can produce different shapes. Consequently, this procedure yields a collection that is rich in shape diversity.

To give these models a photorealistic appearance, we pair them with a specially curated material library containing a wide selection of transparent materials (like glass and plastic) and highly reflective ones (like metal and glazed ceramic).

Scene Creation. This stage focuses on composing scenes dynamically through physics simulation. we randomly select M assets and initialize their six-degree-of-freedom (6-DOF) poses and scales within a predefined environment, such as a container or tabletop, as shown in the topright of Fig. 3, We then employ Blender's integrated physics engine to simulate the objects as they fall and collide, allowing them to settle into a physically plausible and natural final arrangement.

Camera Sampling & Rendering. To capture diverse and dynamic viewpoints, our camera sampling method generates circular trajectories around the geometric center of objects, incorporating sinusoidal perturbations of varying amplitudes. We then utilize Blender's ray tracing engine, Cycles, to perform physically accurate lighting calculations and material rendering. This process precisely simulates complex light transport phenomena, including propagation, refraction, and reflection within transparent materials. As a final step, we use the NVIDIA OptiX-Denoiser to optimize image quality.

The output of this pipeline is **TransPhy3D**, a novel video dataset comprising **11,000** unique scenes. With each scene rendered as a 120-frame video, the dataset contains a total of **1,320,000** frames, sourced from both our parametric and

static asset collections.

B. Preliminaries of Video Diffusion Model

This work builds upon WAN [14], which is comprised of three primary components: a VAE, a diffusion transformer consisting of multiple DiT blocks, and a text encoder. The VAE compresses input videos into latent space and decodes predicted latents back to image space. The text encoder encodes text prompts into embeddings. The diffusion transformer predicts velocity given noisy latents and text embeddings.

WAN leverages the flow matching framework [47] to model a unified denoising diffusion process. During training, given an image or video latent $\mathbf{x}_1$, a random noise $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and a timestep $\mathbf{t} \sim \mathcal{U}(0, 1)$, an intermediate latent $\mathbf{x}_t$ serving as training input is obtained by:

$$\mathbf{x}_t = \mathbf{t}\mathbf{x}_1 + (1 - \mathbf{t})\mathbf{x}_0. \tag{1}$$

The ground truth velocity $\mathbf{v}_t$ is obtained by:

$$\mathbf{v_t} = \frac{\mathbf{dx_t}}{\mathbf{dt}} = \mathbf{x}_1 - \mathbf{x}_0. \tag{2}$$

The loss function is MSE between the output of a velocity predictor $\mathbf{u}$ and $\mathbf{v}_t$:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, \mathbf{c}_{\text{txt}}, \mathbf{t}} \left\| \mathbf{u}\big(\mathbf{x_t}, \mathbf{c}_{\text{txt}}, \mathbf{t}\big) - \mathbf{v}_t \right\|^2, \tag{3}$$

where $\mathbf{c}_{\text{txt}}$ is the text embedding.

C. Training Strategy

As illustrated in Fig. 2, to enhance the training efficiency and alleviate the rendering burden of rendering process. We propose to co-train synthetic image and video data (TransPhy3D).

We first sample a constant number F using:

$$\begin{aligned} F &= 4N + 1 \\ N &\sim \mathcal{U}(0, 5). \end{aligned} \tag{4}$$

F indicates the frame number for the video in this batch of data. Afterwards, if F is equal to 1, we sample a batch of paired data consisting of RGB and depth videos from both video and image datasets (with the video containing only one frame); otherwise, we sample only from video datasets.

Afterwards, the overall pipeline of DKT is presented in Fig. 2, the depth video in one pair is converted to disparity. Both RGB and depth videos are normalized to $[-1, 1]$ to match the VAE training space, then encoded by the VAE into latents $\mathbf{x}_1^c$ and $\mathbf{x}_1^d$.

The depth latent $\mathbf{x}_1^d$ is transformed by Eq. 1 into the intermediate latent $\mathbf{x}_\mathbf{t}^d$. The input to the DiT blocks is then obtained by concatenating $\mathbf{x}_\mathbf{t}^d$ and $\mathbf{x}_1^c$ along the channel dimension. The training loss is defined as the difference between the DiT output and the ground-truth velocity $\mathbf{v}_t^d$ (constructed by Eq. 2), and is computed as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1^d, \mathbf{x}_1^c, \mathbf{c}_{\text{txt}}, \mathbf{t}} \left\| \mathbf{u}\big(\mathbf{Concat}(\mathbf{x}_\mathbf{t}^d, \mathbf{x}_\mathbf{1}^c), \mathbf{c}_{\text{txt}}, \mathbf{t}\big) - \mathbf{v}_t^d \right\|^2, \tag{5}$$

where $\mathbf{c}_{\text{txt}}$ is the text embedding, and **Concat** denotes the concatenation operation along the channel dimension.

All model components remain frozen except for a small set of trainable LoRA [44] parameters in the DiT, which learn low-rank weight adaptations.

D. Implementation

We trained our model with a learning rate of $1e-5$ using AdamW [52] and a batch size of 8. The model is trained using synthetic image datasets, including HISS [4], DREDS [15], and ClearGrasp [16], along with our newly introduced video synthetic dataset, TransPhy3D. Following the training strategy of WAN, all datasets are resized to 832×480 for model training. The number of training iterations is 70K, which takes 8 Nvidia H100 GPUs for two days. Except for specific explanations, the denoising step is set to 5 for inference. By following the inference strategy in [10], we achieve arbitrary-length video inference by splitting the input into overlapping segments. Consecutive segments are then stitched together using a complementary weight applied to the overlapping regions. For more details, please refer to [10].

IV. EXPERIMENT

A. Evaluation Metrics

Following the practice of evaluating the temporal consistency of depth maps[10], the predictions are aligned with the ground truth using a global scale and shift. The following metrics are then calculated: $\delta_{1.05}$, $\delta_{1.10}$, $\delta_{1.25}$, and REL, which are expressed as percentages, along with RMSE, measured in centimeters.

B. Evaluation Datasets

DKT is evaluated using the following real-world and synthetic datasets under a **zero-shot** setting to demonstrate its generalization, robustness, and potential effects on the robotic community.

ClearPose: ClearPose [5] is a real-world RGB-D benchmark for transparent and translucent objects. This testset consists of 27 real-world scenes, including different backgrounds, heavy occlusions, objects in translucent and opaque covers, non-planar surfaces, and even scenes filled with liquid. Each scene is captured in a long video using the RealSense L515 depth camera. The foreground ground truth depth of transparent objects is recovered using object CAD models. The background region is masked by a threshold of $[0.3, 1.5]$ during evaluation. This dataset also provides normal annotation translated from depth map.

DREDS-STD: DREDS-STD [15] is a real-world dataset comprising specular, transparent, and diffuse objects. It is divided into two subsets, CatKnown and CatNovel, based on the commonality of the objects. The former comprises 12 scenes, while the latter contains 5 scenes.

TransPhy3D-Test: We employ a new transparent model [53] to render a synthetic test set using the pipeline we introduce. This dataset consists of 28 scenes.

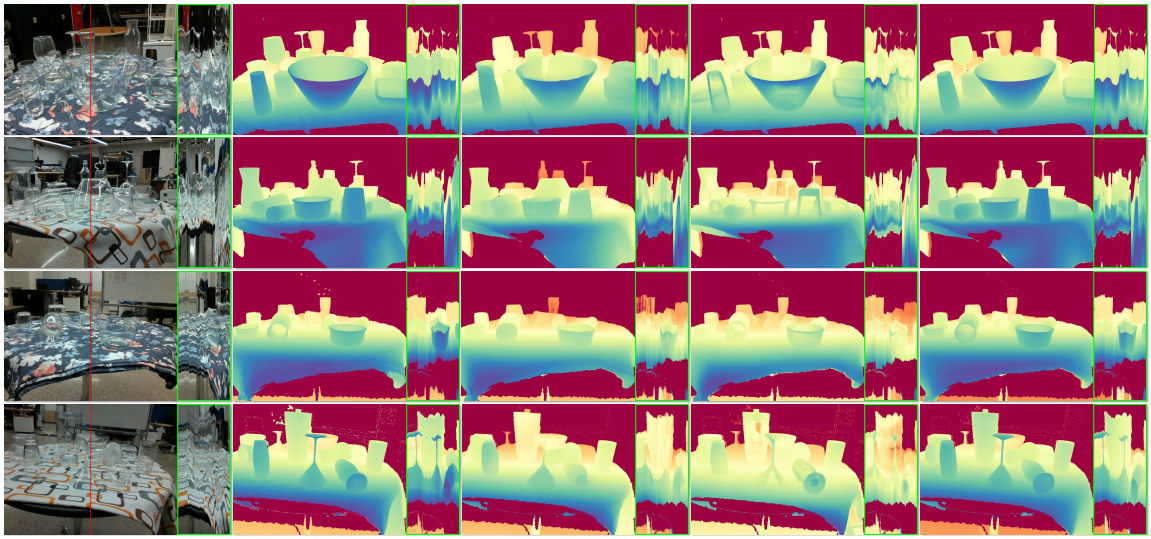

Fig. 4. **Qualitative comparison on the ClearPose [5].** For better visualizing the temporal quality, we show the temporal profiles of each result in green boxes, by slicing the depth values along the time axis at the red line positions.

TABLE I

QUANTITATIVE COMPARISON FOR VIDEO DEPTH ESTIMATION ON CLEARPOSE AND TRANSPHY3D-TEST. **BEST** AND SECOND BEST ARE HIGHLIGHED.

Methods	ClearPose [5]						TransPhy3D-Test					
	REL ↓	RMSE ↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑	Rank ↓	REL ↓	RMSE ↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑	Rank ↓
Depth4ToM [48] (ICCV23)	12.38	14.02	28.74	51.39	85.94	4.0	18.01	720.13	31.54	56.56	85.98	3.8
DAv2 [7] (NeurIPS24)	10.85	**12.21**	32.21	56.37	89.94	**1.8**	14.02	74.86	31.36	53.12	81.27	4.0
Marigold-E2E-FT [49] (WACV25)	16.44	16.65	16.99	33.85	74.24	7.0	23.58	42.63	11.42	27.49	62.86	5.4
MoGe [50] (CVPR25)	13.13	13.40	24.09	45.08	84.29	4.6	31.91	136.74	19.29	32.24	50.98	5.8
VGGT [51] (CVPR25)	15.38	15.68	19.93	38.33	76.89	6.0	32.30	49.82	13.64	30.13	55.25	5.8
DepthCrafter [10] (CVPR25)	11.32	12.34	31.92	55.46	88.59	2.8	11.32	**12.34**	31.92	55.46	88.59	2.0
DKT (Ours)	**9.72**	14.58	**38.17**	**65.50**	**93.04**	**1.8**	**2.96**	19.50	**87.17**	**97.09**	**98.56**	**1.2**

TABLE II

QUANTITATIVE COMPARISON FOR VIDEO DEPTH ESTIMATION ON THE DREDS DATASETS.

Methods	DREDS-STD-CatKnown [15]						DREDS-STD-CatNovel [15]					
	REL ↓	RMSE ↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑	Rank↓	REL ↓	RMSE ↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑	Rank↓
Depth4ToM [48] (ICCV23)	6.92	6.60	44.23	74.15	98.19	5.6	7.21	5.77	43.45	71.13	98.04	5.6
DAv2 [7] (NeurIPS24)	6.94	6.58	44.46	73.66	98.32	5.4	7.41	5.76	41.16	69.77	98.05	6.2
Marigold-E2E-FT [49] (WACV25)	6.07	5.81	49.07	79.15	99.36	3.2	7.06	5.55	42.81	71.51	98.71	4.4
MoGe [50] (CVPR25)	6.95	5.78	47.03	74.44	97.46	4.8	6.07	**4.32**	50.25	78.95	99.31	2.0
VGGT [51] (CVPR25)	5.74	5.14	51.14	80.94	99.79	2.0	6.10	4.74	48.93	78.56	99.53	2.8
DepthCrafter [10] (CVPR25)	7.06	6.41	41.45	72.32	98.68	6.2	7.41	5.54	38.44	70.23	98.51	5.6
DKT (ours)	**5.30**	**4.96**	**53.86**	**84.93**	**99.89**	**1.0**	**5.71**	4.66	**52.12**	**79.51**	**99.84**	**1.2**

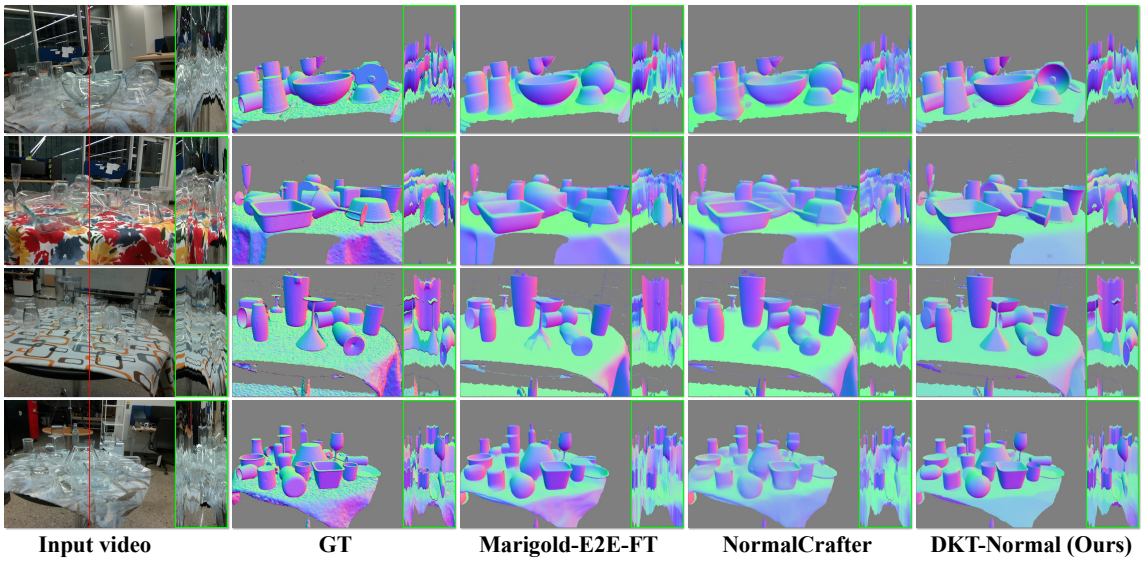

Fig. 5. **Qualitative comparison for video normal estimation on ClearPose [5].**

TABLE III

ABLATION FOR TRAINING STRATEGY IN CLEARPOSE. × INDICATES NAIVE FINETUNING.

Model Size	LoRA	REL ↓	RMSE ↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
1.3B	×	11.86	26.54	30.48	54.03	88.30
1.3B	✓	11.17	17.45	33.16	58.02	90.65
14B	✓	**9.72**	**14.58**	**38.17**	**65.50**	**93.04**

C. State-of-the-art Comparision

A comparison is conducted with available foundation depth estimation methods, including the image depth estimation method Depth-Anything-V2 [7], MoGe [50], VGGT [51], Marigold-E2E-FT [49], Depth4ToM [48] and the video depth estimation method DepthCrafter [10].

As shown in Tab. I and II, DKT sets a new SOTA across three real-world datasets and one synthetic dataset. The performance gap peaks in ClearPose and TransPhy3D, both of which exclusively involve transparent and highly reflective objects. Specifically, we outperform the second-best method by scores of 5.69, 9.13, and 3.1 for $\delta_{1.05}$, $\delta_{1.10}$, and $\delta_{1.25}$ in ClearPose, and 55.25, 40.53, and 9.97 for $\delta_{1.05}$, $\delta_{1.10}$, and $\delta_{1.25}$ in TransPhy3D.

Moreover, this advantage is clearly reflected in Fig. 4. Following [10], we present a temporal profile to better illustrate the temporal consistency of the predicted video depth. DKT not only achieves superior identification of transparent objects in the first frame but also demonstrates optimal temporal consistency.

What accounts for the significant performance gap in the TransPhy3D-Test? We attribute this phenomenon to a characteristic of our rendering data: the camera trajectories during rendering describe a circular path around the object. This feature increases the requirements for inter-frame consistency in the model's predictions, as even a minor error can lead to a significant decline in prediction accuracy after global alignment.

D. Ablation Study

Training Strategies. As illustrated in Tab. III, naive finetuning results in high computational costs and suboptimal performance compared to LoRA fine-tuning. By adopting the LoRA training strategy and scaling up the model size, a significant improvement in perforamnce is achieved.

Inference Steps As illustrated in the upper part of Fig. 6, increasing the number of inference steps does not yield significant performance improvements. Moreover, as shown in the lower part of Fig. 6, fewer steps result in inaccurate predictions, while more inference steps lead to the loss of important details. To balance performance and inference efficiency, we set 5 as the default number of inference steps.

Computational Efficiency To assess inference efficiency fairly, we reevaluated DKT, the baseline DAv2-Large [7], and DepthCrafter [10] on the Nvidia L20. The results, shown in Tab. IV, indicate that DKT-1.3B achieves the highest efficiency, with a remarkable inference time of only 167.48ms per frame, surpassing DAv2 by 110.27ms. Moreover, the peak GPU memory occupancy of DKT-1.3B is only 11.19 GB, which is acceptable for most robot computational platforms.

Inference Steps	REL ↓	RMSE ↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
30	9.87	**13.88**	36.58	63.98	92.56
20	9.81	13.94	36.90	64.42	92.71
15	9.76	14.04	37.21	64.78	92.80
10	9.74	14.27	37.50	65.19	92.91
5	**9.72**	14.58	**38.17**	**65.50**	**93.04**
1	10.00	15.07	37.64	63.65	92.65

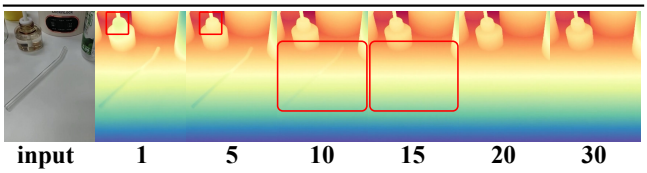

Fig. 6. **Effect of the number of inference steps. Fewer steps lead to inaccurate predictions, whereas more steps cause loss of fine details in small objects. Five steps provide the best trade-off between quality and efficiency.**

TABLE IV

INFERENCE TIME PER FRAME (MS) AT A RESOLUTION OF 832 × 480.

Method	Encoding	Denoising	Decoding	All
DAv2 [7]	N/A	N/A	N/A	277.75
DepthCrafter [10]	141.85	240.01	183.69	565.55
DKT-14B	46.53	297.11	68.07	411.71
DKT-1.3B	46.53	52.88	68.07	167.48

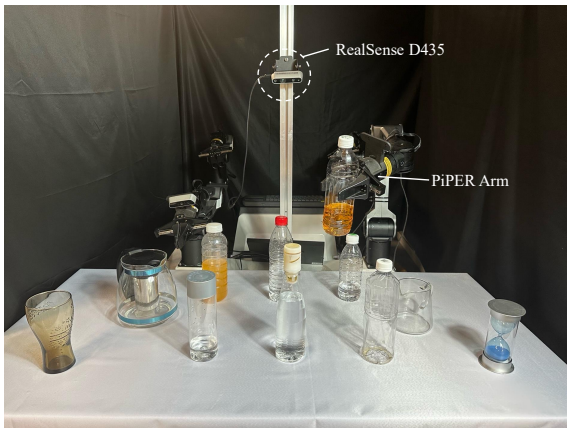

Fig. 7. **Real-World Setup.** We use the Cobot Magic system, which integrates the PiPER Arm and RealSense D435 for our grasping tasks.

Video Normal Estimation To further validate the efficacy of our training strategy, DKT-Normal-14B is introduced following the same training strategy of DKT-14B. As demonstrated in Tab. V, DKT-Normal-14B significantly outperforms the previous SOTA video normal estimation method NormalCrafter [43] and the previous SOTA image estimation method Marigold-E2E-FT [49] by a substantial margin on the zero-shot dataset ClearPose, using the same metrics as NormalCrafter. Moreover, DKT-Normal-14B achieves the sharpest normal prediction and the highest temporal consistency, as illustrated in Fig. 5.

E. Real-world Grasping Experiments

Experimental Setup. *1) Hardware System:* As shown in Fig. 7, our real-world robotic manipulation environment con-

TABLE V

VIDEO NORMAL ESTIMATION ON THE CLEARPOSE.

	mean↓	med↓	11.25° ↑	22.5° ↑	30° ↑
NormalCrafter [43] (ICCV25)	27.08	20.29	26.10	55.37	68.81
Marigold-E2E-FT [49] (WACV25)	27.08	19.40	29.78	57.22	69.30
DKT-Normal-14B	**26.03**	**18.59**	**30.06**	**59.63**	**70.98**

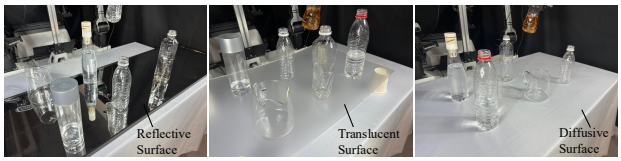

Fig. 8. **Demonstration of Surface Types.**

sists of two PiPER ARM manipulators used to grasp objects and a fixed-view Realsense D435 camera to provide RGB observations. *2) Environment Arrangement:* As illustrated in Fig. 8, we set up three types of tabletop backgrounds: reflective, translucent, and diffusive surfaces. Various objects with specular, transparent, and diffusive properties are placed on the table. This setup allows for a comprehensive evaluation of the effectiveness and robustness of different methods under complex real-world scenarios.

Deployment Pipeline. Initially, an RGB image is captured by the D435 camera and processed with various relative depth estimation models, including DAv2-Large, DepthCrafter, and DKT-1.3B, to generate relative depth maps. These are then rescaled to metric depth using AprilTag [54]. Subsequently, the RGB image and metric depth are input into AnyGrasp [55] to generate a 7-DoF grasp pose (end-effector pose + gripper width). Finally, CuRobo [56] is utilized to plan an executable trajectory, which is executed by the PiPER ARM.

Results and Analysis. As demonstrated in Tab. VI, DKT consistently outperforms all baseline across all settings by a significant margin. For the perception results and the grasping video, please refer to the appendix.

V. CONCLUSIONS

In this work, we dive into the challenging problem of video depth and normal estimation for transparent and highly reflective scenarios, which is crucial for robotic perception. Our contributions can be summarized in three parts. First, the first video dataset for transparent and highly reflective objects is introduced, accompanied by diverse transparent asset categories and an infinite variety of 3D shapes. Second, DKT is introduced, finetuned from a video diffusion model using a LoRA training strategy. Finally, we demonstrate DKT's performance in comprehensive benchmarks, including synthetic and real-world datasets, and DKT sets a new SOTA across all of them in video depth and normal estimation.

TABLE VI

GRASPING SUCCESS RATE OF DIFFERENT DEPTH ESTIMATORS ON TRANSLUCENT, REFLECTIVE, DIFFUSIVE SURFACE, RESPECTIVELY.

Method	Translucent	Reflective	Diffusive	Mean
RAW	0.47	0.18	0.56	0.384
DAv2	0.60	0.27	0.56	0.46
DepthCrafter	0.67	0.23	0.625	0.48
DKT-1.3B	**0.80**	**0.59**	**0.81**	**0.73**

REFERENCES

[1] M. Grinvald, F. Tombari, R. Siegwart, and J. Nieto, "Tsdf++: A multi-object formulation for dynamic object tracking and reconstruction," in *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 14192–14198, IEEE, 2021.

[2] W. Cui, C. Zhao, S. Wei, J. Zhang, H. Geng, Y. Chen, H. Li, and H. Wang, "Gapartmanip: A large-scale part-centric dataset for material-agnostic articulated object manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14791–14798, IEEE, 2025.

[3] J. Liu, H. Ma, Y. Guo, Y. Zhao, C. Zhang, W. Sui, and W. Zou, "Monocular depth estimation and segmentation for transparent object with iterative semantic and geometric fusion," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11162–11168, 2025.

[4] S. Wei, H. Geng, J. Chen, C. Deng, C. Wenbo, C. Zhao, X. Fang, L. Guibas, and H. Wang, "D^3 roma: Disparity diffusion-based depth sensing for material-agnostic robotic manipulation," in *ECCV 2024 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild*, 2024.

[5] X. Chen, H. Zhang, Z. Yu, A. Opipari, and O. Chadwicke Jenkins, "Clearpose: Large-scale transparent object dataset and benchmark," in *European conference on computer vision*, pp. 381–396, Springer, 2022.

[6] J. Shi, A. Yong, Y. Jin, D. Li, H. Niu, Z. Jin, and H. Wang, "Asgrasp: Generalizable transparent object reconstruction and 6-dof grasp detection from rgb-d active stereo camera," in *2024 IEEE international conference on robotics and automation (ICRA)*, pp. 5441–5447, IEEE, 2024.

[7] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21875–21911, 2024.

[8] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

[11] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[12] H. Geng, S. Wei, C. Deng, B. Shen, H. Wang, and L. Guibas, "Sage: Bridging semantic and actionable parts for generalizable articulated-object manipulation under language instructions," *arXiv preprint arXiv:2312.01307*, vol. 2, 2023.

[13] DeepMind, "Veo: A text-to-video generation system," Technical Report Veo-3 Tech Report, Google DeepMind, 2024. Describes the components of Veo 3, including the diffusion-based audio+video model, training data, and safety evaluations.

[14] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, M. Feng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu, Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z.-F. Wu, and Z. Liu, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

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

[16] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "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.

[17] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "Fastdepth: Fast

monocular depth estimation on embedded systems," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6101–6108, IEEE, 2019.

[18] J. Okae, B. Li, J. Du, and Y. Hu, "Robust scale-aware stereo matching network," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 2, pp. 244–253, 2021.

[19] B. Li, Y. Sun, Z. Liang, D. Du, Z. Zhang, X. Wang, Y. Wang, X. Jin, and W. Zeng, "Bridging stereo geometry and bev representation with reliable mutual interaction for semantic scene completion," *arXiv preprint arXiv:2303.13959*, 2023.

[20] H. Li, Y. Ma, Y. Gu, K. Hu, Y. Liu, and X. Zuo, "Radarcam-depth: Radar-camera fusion for depth estimation with learned metric scale," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10665–10672, IEEE, 2024.

[21] Y. Liao, L. Huang, Y. Wang, S. Kodagoda, Y. Yu, and Y. Liu, "Parse geometry from a line: Monocular depth estimation with partial laser observation," in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 5059–5066, IEEE, 2017.

[22] S. S. Shivakumar, K. Mohta, B. Pfrommer, V. Kumar, and C. J. Taylor, "Real time dense depth estimation by fusing stereo with sparse depth measurements," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6482–6488, IEEE, 2019.

[23] S. Pillai, R. Ambruş, and A. Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9250–9256, IEEE, 2019.

[24] C. Wang, Y. Qin, Z. Kang, N. Ma, and R. Zhang, "Toward accurate camera-based 3d object detection via cascade depth estimation and calibration," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2006–2012, IEEE, 2024.

[25] L. Ebner, G. Billings, and S. Williams, "Metrically scaled monocular depth estimation through sparse priors for underwater robots," in *2024 IEEE international conference on robotics and automation (ICRA)*, pp. 3751–3757, IEEE, 2024.

[26] A. Li, A. Hu, W. Xi, W. Yu, and D. Zou, "Stereo-lidar depth estimation with deformable propagation and learned disparity-depth conversion," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2729–2736, IEEE, 2024.

[27] B. Li, Y. Sun, J. Dong, Z. Zhu, J. Liu, X. Jin, and W. Zeng, "One at a time: Progressive multi-step volumetric probability learning for reliable 3d scene perception," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 3028–3036, 2024.

[28] B. Ke, D. Narnhofer, S. Huang, L. Ke, T. Peters, K. Fragkiadaki, A. Obukhov, and K. Schindler, "Video depth without video models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7233–7243, 2025.

[29] G. Chou, W. Xian, G. Yang, M. Abdelfattah, B. Hariharan, N. Snavely, N. Yu, and P. Debevec, "Flashdepth: Real-time streaming video depth estimation at 2k resolution," *arXiv preprint arXiv:2504.07093*, 2025.

[30] H. Ni, B. Egger, S. Lohit, A. Cherian, Y. Wang, T. Koike-Akino, S. X. Huang, and T. K. Marks, "Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9015–9025, 2024.

[31] B. Li, J. Guo, H. Liu, Y. Zou, Y. Ding, X. Chen, H. Zhu, F. Tan, C. Zhang, T. Wang, *et al.*, "Uniscene: Unified occupancy-centric driving scene generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11971–11981, 2025.

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

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

[34] A. Cheng, Z. Yang, H. Zhu, and K. Mao, "Gam-depth: self-supervised indoor depth estimation leveraging a gradient-aware mask and semantic constraints," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5367–5374, IEEE, 2024.

[35] S. Dai, X. Lou, P. Nilsson, S. Thakar, C. Meeker, A. Gordon, X. Kong, J. Zhang, B. Knoerlein, R. Liu, *et al.*, "Depth estimation through translucent surfaces," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1550–1557, IEEE, 2025.

[36] L. Zhong, Y. Zhang, H. Zhao, A. Chang, W. Xiang, S. Zhang, and L. Zhang, "Seeing through the occluders: Robust monocular 6-dof object pose tracking via model-guided video object segmentation," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5159–5166, 2020.

[37] X. Chen, J. Liu, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Nerrf: 3d reconstruction and view synthesis for transparent and specular objects with neural refractive-reflective fields," *arXiv preprint arXiv:2309.13039*, 2023.

[38] C. Zhong, Y. Zheng, Y. Zheng, H. Zhao, L. Yi, X. Mu, L. Wang, P. Li, G. Zhou, C. Yang, *et al.*, "3d implicit transporter for temporally consistent keypoint discovery," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3869–3880, 2023.

[39] Y. Shuai, R. Yu, Y. Chen, Z. Jiang, X. Song, N. Wang, J. Zheng, J. Ma, M. Yang, Z. Wang, *et al.*, "Pugs: Zero-shot physical understanding with gaussian splatting," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4478–4485, IEEE, 2025.

[40] Q. Sun, C. Shu, S. Zhou, R. Cheng, Y. Wei, Z. Yu, D. Yang, S. Han, and Y. Chun, "Gsrender: Deduplicated occupancy prediction via weakly supervised 3d gaussian splatting," *arXiv preprint arXiv:2412.14579*, 2024.

[41] L. Zhang and M. Agrawala, "Transparent image layer diffusion using latent transparency," *arXiv preprint arXiv:2402.17113*, 2024.

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

[43] Y. Bin, W. Hu, H. Wang, X. Chen, and B. Wang, "Normalcrafter: Learning temporally consistent normals from video diffusion priors," *arXiv preprint arXiv:2504.11427*, 2025.

[44] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[45] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[46] Y. H. Kim, S. Kim, Y. Lee, and F. C. Park, "T^2sqnet: A recognition model for manipulating partially observed transparent tableware objects," in *CORL*, 2024.

[47] Y. Lipman, M. Havasi, P. Holderrieth, N. Shaul, M. Le, B. Karrer, R. T. Chen, D. Lopez-Paz, H. Ben-Hamu, and I. Gat, "Flow matching guide and code," *arXiv preprint arXiv:2412.06264*, 2024.

[48] A. Costanzino, P. Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano, "Learning depth estimation for transparent and mirror surfaces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9244–9255, 2023.

[49] G. Martin Garcia, K. Abou Zeid, C. Schmidt, D. de Geus, A. Hermans, and B. Leibe, "Fine-tuning image-conditional diffusion models is easier than you think," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

[50] R. Wang, S. Xu, C. Dai, J. Xiang, Y. Deng, X. Tong, and J. Yang, "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, 2025.

[51] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

[52] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[53] J. Kim, M.-H. Jeon, S. Jung, W. Yang, M. Jung, J. Shin, and A. Kim, "Transpose: Large-scale multispectral dataset for transparent object," *The International Journal of Robotics Research*, vol. 43, no. 6, pp. 731–738, 2024.

[54] J. Wang and E. Olson, "Apriltag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4193–4198, IEEE, 2016.

[55] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[56] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. Van Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos, *et al.*, "curobo: Parallelized collision-free minimum-jerk robot motion generation," *arXiv preprint arXiv:2310.17274*, 2023.