

3D Scene Understanding Through Local Random Access Sequence Modeling

Wanhee Lee^{1*} Klemen Kotar^{1*} Rahul Mysore Venkatesh^{1*} Jared Watrous^{1*} Honglin Chen^{2*}

Khai Loong Aw¹ Daniel L. K. Yamins¹

¹Stanford University, ²OpenAI

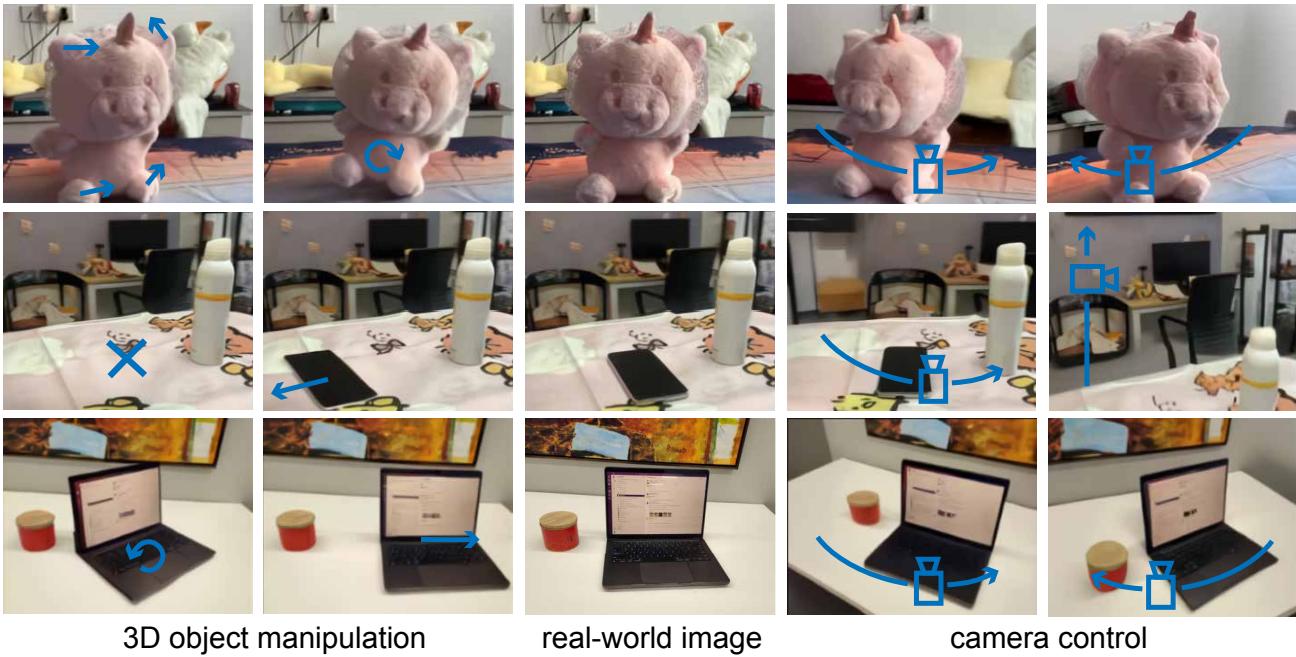


Figure 1. Our model (**LRAS**) enables sophisticated object manipulation (left) and camera control (right) on real-world images (middle). The generation exhibits understanding of 3D scene structure and properties of the visual scenes, such as lighting, shadows, continuity, occlusions, and amodal completion.

Abstract

3D scene understanding from single images is a pivotal problem in computer vision with numerous downstream applications in graphics, augmented reality, and robotics. While diffusion-based modeling approaches have shown promise, they often struggle to maintain object and scene consistency, especially in complex real-world scenarios. To address these limitations, we propose an autoregressive generative approach called Local Random Access Sequence (**LRAS**) modeling, which uses local patch quantization and randomly ordered sequence generation. By utilizing optical flow as an intermediate representation for 3D scene editing, our experiments demonstrate that **LRAS** achieves state-of-the-art novel view synthesis and 3D object manipulation ca-

pabilities. Furthermore, we show that our framework naturally extends to self-supervised depth estimation through a simple modification of the sequence design. By achieving strong performance on multiple 3D scene understanding tasks, **LRAS** provides a unified and effective framework for building the next generation of 3D vision models.[†]

1. Introduction

Understanding 3D scenes from a single image remains a fundamental yet unsolved challenge in computer vision, and a necessary prerequisite for many robotics tasks. In this work, we study 3D scene understanding in the context of three tasks: a) Novel view synthesis – understanding how the scene changes when the camera moves, b) 3D object manipulation – which tests the model’s ability to predict object appearance changes under rigid transformations, and c) Depth estimation – asking how well the model perceives the 2.5D structure of visible regions.

*Equal contribution to this work.

[†]Project website at: https://neuroailab.github.io/projects/lras_3d/

The most dominant approaches to solving novel view synthesis and object manipulation tasks have been fine-tuning large diffusion models pre-trained on text to image or video generation [31, 50, 59]. Although these methods demonstrate strong capabilities and produce photorealistic images, they have certain key limitations. We benchmark the performance of these methods on in-the-wild data and find that they often fail to preserve object identity, shift global lighting, and provide imprecise control over camera and object motion. Additionally, several of these models [31, 59] rely on separate off-the-shelf supervised depth estimation models as part of their editing pipeline, sidestepping the question of how depth estimation can emerge in these large pre-trained models.

An alternative approach is to use LLM-inspired [3] autoregressive next-token prediction for generative image modeling. Recently, such models [45, 57] have emerged as a strong alternative to diffusion, outperforming diffusion baselines on image generation tasks. Unlike diffusion models, which build images from the “top down”, by turning noise into rough outlines of shapes and then filling in the detailed textures, autoregressive models build up an image from the “bottom up”, predicting the image patch by patch. However, in practice, most autoregressive models predict sequences of globally encoded tokens. In addition, these models predict sequences in raster order, allowing “top left” tokens to have greater causal control of the predicted image, which leads to inferior generation [24].

Our model, **LRAS** (**L**ocal **R**andom **A**ccess **S**equence Model) addresses these shortcomings in autoregressive image modeling and gets its name from the two key innovations: a) **L**ocal patch representations and b) **R**andom order decoding. In our model we predict a sequence of local patch representations which is more in line with the standard autoregressive next token prediction paradigm in LLMs. Next, we introduce architectural innovations that equip the model to decode the image in spatially random order by predicting a sequence of (pointer, contents) representations – where the pointer indicates the spatial location at which the contents should be placed.

We take this model, and apply it to 3D understanding tasks by using optical flow intermediates. First, using **LRAS_{RGB}** we learn to predict RGB images conditioned on an input frame and an optical flow map. We demonstrate that this model possesses emergent 3D scene editing capabilities such as NVS and 3D object manipulation. Further, we find that the **LRAS** framework is flexible and can also be used as a camera conditioned flow predictor (**LRAS_{FLOW}**) which we use to extract 2.5D depth, addressing the challenge of emergent self-supervised depth extraction from large pre-trained models. To train our model, we crawl a dataset of 7k hours of high-quality, diverse internet videos called Big Video Dataset (**BVD**). We demonstrate that **BVD** can be used

to train powerful generative models.

We provide empirical evidence showing the effectiveness of our approach across multiple 3D vision tasks. For novel view synthesis, our method achieves state-of-the-art performance on both object-centric and scene-level datasets. Further, to assess object manipulation capabilities, we introduce **3DEditBench**, a new real-world object editing benchmark. Our evaluation demonstrates that our model outperforms competing object manipulation methods on real-world data. Notably, **LRAS** exhibits a significant advantage over diffusion models in preserving scene structure, object identity, and global illumination during 3D editing tasks. We also find that our model achieves state-of-the-art self-supervised depth estimation results on standard benchmarks on both static and dynamic objects, which has previously been hard to achieve with geometric consistency methods [44, 63]. In this way, **LRAS** emerges as a foundational model of 3D vision with a wide range of capabilities.

2. Related Works

Novel View Synthesis (NVS) has been widely studied as a fundamental task in 3D vision. Regression-based approaches [5, 22, 38, 56] perform well for view interpolation but struggle with single-image-to-3D synthesis, producing blurry results due to uncertainty in occluded regions. This limitation has driven a shift toward generative models, particularly diffusion-based methods, which enable high-quality and diverse NVS. Zero-1-to-3 [26], trained on large-scale synthetic datasets [4, 6], predicts novel views from a single image using implicit camera modeling. ZeroNVS [39] integrates Zero-1-to-3’s approach with a score distillation sampling framework [34], and extends the application to real-world scenes. Other approaches, such as MotionC-ctrl [50], inject camera embeddings to guide video diffusion without explicit 3D representations. Recently, ViewCrafter [59] utilized point-cloud rendering using DUS3R [48] for improved performance with better camera motion control. In this work, we explore autoregressive sequence modeling for the NVS problem as an alternative to diffusion-based approaches to overcome the limitations of previous works.

3D Object Manipulation While NVS focuses on generating novel views of the input scene, object manipulation refers to the task of transforming objects in the scene while keeping the camera fixed. Drag-based image editing methods [41, 50, 52, 55] aim to solve this problem by parameterizing object transforms as 2D motion vectors which are then used as conditioning to fine-tune stable diffusion (SD) [37]. These methods can be naturally extended to more complex 3D transforms by incorporating depth information into the drag vectors [49]. Another class of models [20, 31], performs 3D object manipulations by editing input depth maps according to the desired object transform and utilizing a depth-conditioned diffusion model to generate the edited

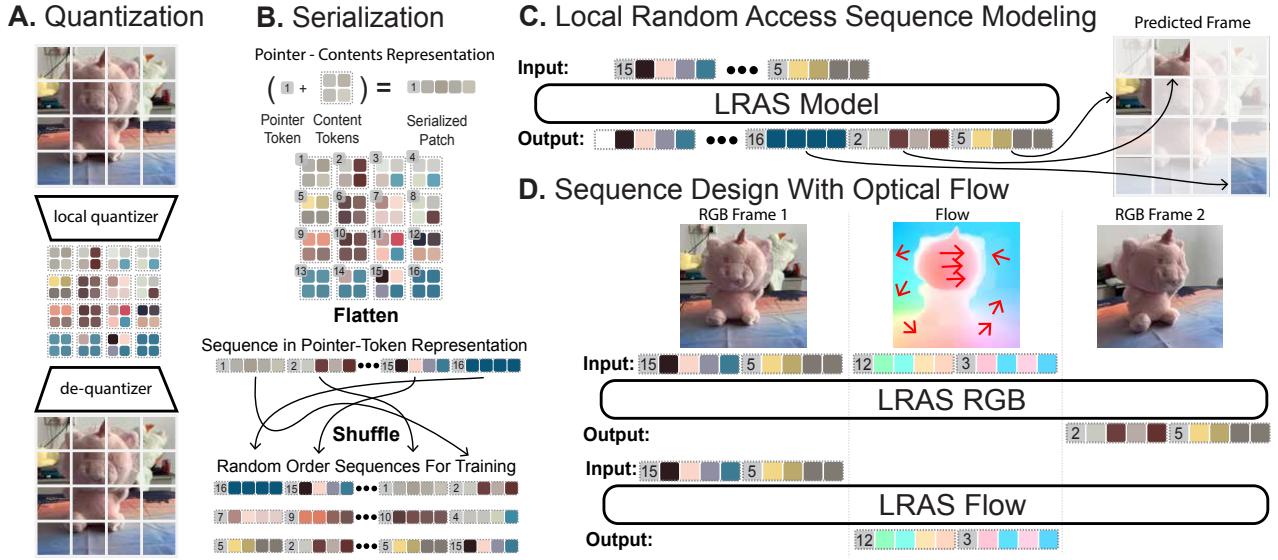


Figure 2. LRAS Architecture. **A. Quantization:** We train a small, patch local, convolutional autoencoder with a 16 bit LFQ codebook. **B. Serialization:** We serialize the codes into sequences using the pointer-content representation, which allows us to arbitrarily order the patches during training and generation. **C. Local Random Access Sequence Modeling:** We train an LLM-like autoregressive transformer to predict the contents of the next patch, shuffled in random order. **D. Sequence Design With Optical Flow:** We design sequences of tokens that contain optical flow intermediates, to provide robust control over the generation. We train two models: **LRAS_{RGB}**, which is conditioned on a source RGB image and an optical flow describing the desired transformation to predict the next frame, and **LRAS_{FLOW}**, which is conditioned on a source RGB image to predict a plausible optical flow field.

image. However, these methods heavily rely on inverting the input image into the SD latent space, which often fails on real-world images [29]. In contrast, our model **LRAS** is an autoregressive sequence model trained from scratch on internet videos, which does not rely on SD and is free from inversion processes. We find that it generalizes better to real-world images and makes more accurate edits compared to prior approaches.

Concurrent Work Recently, several works that have shown that motion-conditioned diffusion models can be used to perform sophisticated image manipulations. [12, 17, 21] trains a spatio-temporal trajectory-conditioned control net on top of a large video diffusion model [1]. The model demonstrates emergent capabilities such as object and camera control and drag-based image editing and motion transfer. Another set of recent work [11, 15, 61] uses 3D point trajectories providing more powerful control over image generation. **LRAS** is the first model that explores the idea of using motion conditioning to train autoregressive image generation models. We find that our model demonstrates strong performance on NVS and object manipulation tasks compared to its diffusion model counterparts.

3. Method

3.1. LRAS Architecture

The Local Random Access Sequence Model is an autoregressive transformer with two key properties - locality and

random access. **Locality** is achieved by utilizing an image quantizer which produces a grid of codes that only contain information from their corresponding patch of the input image (as illustrated in Figure 2A). This way each token is independent of all others in the sequence - a property which is naturally present in most text tokenization schemes employed by LLMs, but is missing from modern image quantizers such as VQ-GAN [8] or the COSMOS tokenizer [9]. We hypothesize that, as with language, this gives the sequence model stronger downstream compositional abilities, as it learns to model objects as groups of individual tokens without any global dependencies on other objects or the scene. **Random Access** is achieved by the addition of pointer tokens to the sequence (as illustrated in Figure 2B). Since autoregressive transformers operate over 1D sequences, they have to process information in a serial order. The addition of pointers allows them to arbitrarily jump around the sequence, filling in parts of the data structure in any order. Additionally, the pointers themselves could be predicted, allowing the model to drive the generation order; we leave this exploration for future work.

3.2. Local Patch Quantization

Our tokenizer is a fully convolutional autoencoder with 40M parameters, with a 16 bit LPQ Bottleneck [58] for discretizations, giving us a vocabulary size of 65,536. Our encoder consists of three ResNet [16] style blocks. The first

layer has kernel size 4 and stride of 4, reducing the image to a 64x64 grid feature map, while subsequent encoder layers have kernel sizes and strides of 1. This design enforces that no information is shared between adjacent input patches. After the quantization layer, we apply a convolutional decoder with 6 ResNet style blocks and a kernel size 3 and stride of 1. This allows for some local information sharing between adjacent patches to make the reconstructed image coherent. The model is supervised using only L2 regression loss, and is trained on frames from the Kinetics400 [18] dataset. We train a second encoder, identical in architecture, to quantize optical flow fields with a 32,768-token vocabulary, using RAFT [47] flow from Kinetics400.

3.3. Random Access Through Pointer-Contents Representation

After quantization, an image needs to be serialized into a 1D sequence. Unlike text, which naturally follows this format, images and videos require a certain chosen order. Traditional autoregressive models use a fixed scanning order, but as shown in [24], this is suboptimal. Instead, we allow the model to generate in arbitrary order. While [32] and [23] concurrently achieve this by passing two positional embeddings to each token - one for the current token and one for the next token to generate, we take an alternative approach - the **Pointer Token**. These special tokens guide the model across the entire sequence by allowing it to “jump” to a new location during encoding or generation. Each pointer token is followed by the **Content Tokens**, which contain the actual RGB or Flow information at that location. During training, this allows us to randomly shuffle the order in which images are decoded, and train on only subsets of the image patches - since the image generation problem gets easier the more patches we reveal, and thus the supervision on the latter tokens is less useful. At test time, this allows us to control the order in which we predict the image, as well as only predict parts of the image or perform some of the prediction in parallel.

3.4. Optical Flow Conditioning

While the **LRAS** formulation is fairly general, in this work we focus on applying it to 3D scene understanding, by utilizing optical flow intermediates. As shown concurrently in [14, 31, 40], this formulation allows us to express any physical scene edit in the space of flow fields, yielding precisely conditioned RGB generations with diverse hallucinations. We utilize optical flow as conditioning, and uniquely also a prediction target. We illustrate how this approach naturally fits with autoregressive models, and obtain state-of-the-art results on a number of challenging 3D scene editing tasks. We introduce two models (as shown in figure 2D):

LRAS_{RGB} is a 7B parameter model, which takes as input an RGB frame and a dense flow field (both quantized by

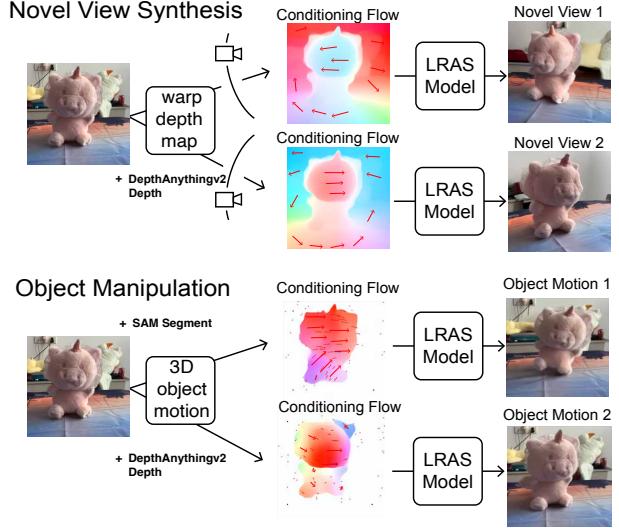


Figure 3. 3D Scene Editing Through Flow Field Manipulation: We perform 3D scene edits by constructing optical flow fields corresponding to the desired transformations - either camera or object motion in 3D.

a local patch quantizer) and predicts the next RGB frame. **LRAS_{FLOW}**, is a 1B parameter model with the same architecture, which utilizes the flow tokens as its target, and is conditioned only on the first frame (and when available in the data, a camera pose change signal). Next, we will describe how we use **LRAS_{RGB}** for 3D scene editing and **LRAS_{FLOW}** for depth extraction.

3.5. Dataset and Training

LRAS was pre-trained on a large dataset containing internet videos, called **BVD** (big video dataset), along with 3D vision datasets including the train splits of ScanNet++ [54], CO3D [36], RealEstate10K [64], MVImgNet [60], DL3DV [25], and EgoExo4D [13] dataset. We used RAFT [47] to compute the optical flow from the videos for the training. Further information on the dataset can be found in the supplementary materials.

Our models were trained in an autoregressive fashion with cross-entropy loss applied on next token prediction. For **LRAS_{RGB}**, only the next frame RGB token targets are supervised. For **LRAS_{FLOW}**, only the flow tokens are supervised. Each model is optimized for 500,000 steps with a batch size of 512.

3.6. Model Inference

Novel View Synthesis can be performed using **LRAS_{RGB}** by conditioning the model on 2D optical flow fields that represent how the pixels move given a desired camera pose change. To generate these flow fields, we use the following steps: a) unproject the depthmap of the input image to obtain a 3D point cloud, b) apply a rigid transformation to the point cloud as per the given camera transformation, c)

Dataset	Model	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
WildRGB-D	MotionCtrl	12.394	0.293	0.404
	ZeroNVS	16.143	0.460	0.283
	ViewCrafter	13.960	0.375	0.290
	LRAS (Ours)	17.748	0.536	0.218
DL3DV	MotionCtrl	12.629	0.261	0.462
	ZeroNVS	15.622	0.403	0.331
	ViewCrafter	16.592	0.430	0.253
	LRAS (Ours)	18.110	0.523	0.328

Table 1. Comparison of metrics for novel view synthesis.

re-project the transformed point cloud and compute the displacement relative to the pixels of the first frame to compute the 2D flow (See Figure 3). **LRAS_{RGB}** generates the edited image given the computed flow map and the input image. As we will describe in the next section, **LRAS_{FLOW}** provides a natural method of extracting depth maps in a self-supervised manner. However, in practice we find that marginally better performance can be achieved using off-the-shelf supervised metric depth estimators such as Depth Anything V2 [53].

3D Object Manipulation can be performed by creating a flow field where the flow on the surface of the object characterizes the 3D transformation to be performed, with the flow of the background set to 0 – conditioning the predictor to move the object, but keep the background fixed. We follow a similar procedure described above to produce flow fields for rigid object transformations and use the SegmentAnything [19] model to suppress the flow of the background regions (See Figure 3).

Depth Extraction and End-to-End NVS Camera conditioned **LRAS_{FLOW}** provides a natural method for extracting depth maps without additional finetuning. We provide in-plane camera motion as input to **LRAS_{FLOW}** and predict the optical flow induced by camera motion. We compute the magnitude of the optical flow to compute the disparity which, when inverted, yields 2.5D depth maps. In practice, we find that a simple upward camera translation is sufficient to generate high-quality depth maps. Additionally, performance can be improved by statistical aggregation over disparity maps generated with different seeds for the same image. These depth maps can be used in conjunction with **LRAS_{RGB}** for end-to-end NVS without relying on off-the-shelf depth estimators.

4. Results

4.1. Novel View Synthesis

Evaluation Details To ensure a fair evaluation of novel view synthesis (NVS) on out-of-distribution datasets, we

selected two benchmarks: WildRGB-D for object-centric NVS and DL3DV for scene-level NVS. For WildRGB-D, we randomly sampled 100 scenes from its evaluation split. For DL3DV, since some models were trained on this dataset, we selected 100 scenes from its recently released 11K subset, which, to the best of our knowledge, was not used to train any of the compared models. From each video, we extracted a 25-frame sequence and used the first frame as the input image and evaluating the generated frames. For quantitative evaluation, we measured PSNR, SSIM, and LPIPS [62]. As baselines, we compared against MotionCtrl, ZeroNVS, and ViewCrafter. Further implementation details are provided in the supplementary materials.

Qualitative and Quantitative Comparisons As shown in Figure 4, our model achieves high-quality novel view reconstruction that maintains object and scene identity. In contrast, MotionCtrl distorts the scene and objects inconsistently. ZeroNVS often suffers from inaccurate 3D reconstruction and artifacts, and its hallucinated regions often become blurry and unrealistic. ViewCrafter may produce visually appealing images, but it frequently changes object appearance and global illumination. Our approach, built on local token-based autoregressive transformers, ensures object and global scene identity remain consistent. Our model also demonstrates robust and precise camera control, a significant advantage over previous methods. Despite our efforts to optimize scene scales, MotionCtrl fails to accurately control camera motion regardless of conditioning, while ZeroNVS faces pose alignment issues when its 3D reconstruction quality is poor. In contrast, our model directly computes pixel correspondences from depth, providing more intuitive and reliable control over camera motion. The key difference in scene alignment between ViewCrafter and our method is that the former optimizes scale in 3D point cloud space, whereas our method performs scale optimization directly in pixel space using optical flow correspondences.

Quantitatively, our model outperforms previous methods in reconstruction quality metrics, as shown in Table 1. It achieves the best overall metrics on WildRGB-D, and the best PSNR and SSIM scores on DL3DV, although ViewCrafter attains a better LPIPS. These results reflect our model’s performance with precise camera control and preservation of scene and object identity.

4.2. 3D Object Manipulation

Baselines We compare to DiffusionHandles [31], which is the closest related work that performs 3D object edits using depth-conditioned diffusion models. Additionally, we also compare to drag-based image editing models such as LightningDrag [41] and DragAnything [52]. Although these methods cannot be directly conditioned on 3D transforms, we find that providing sparse 2D flow vectors (which are part of our dataset’s annotations) can be used to make these

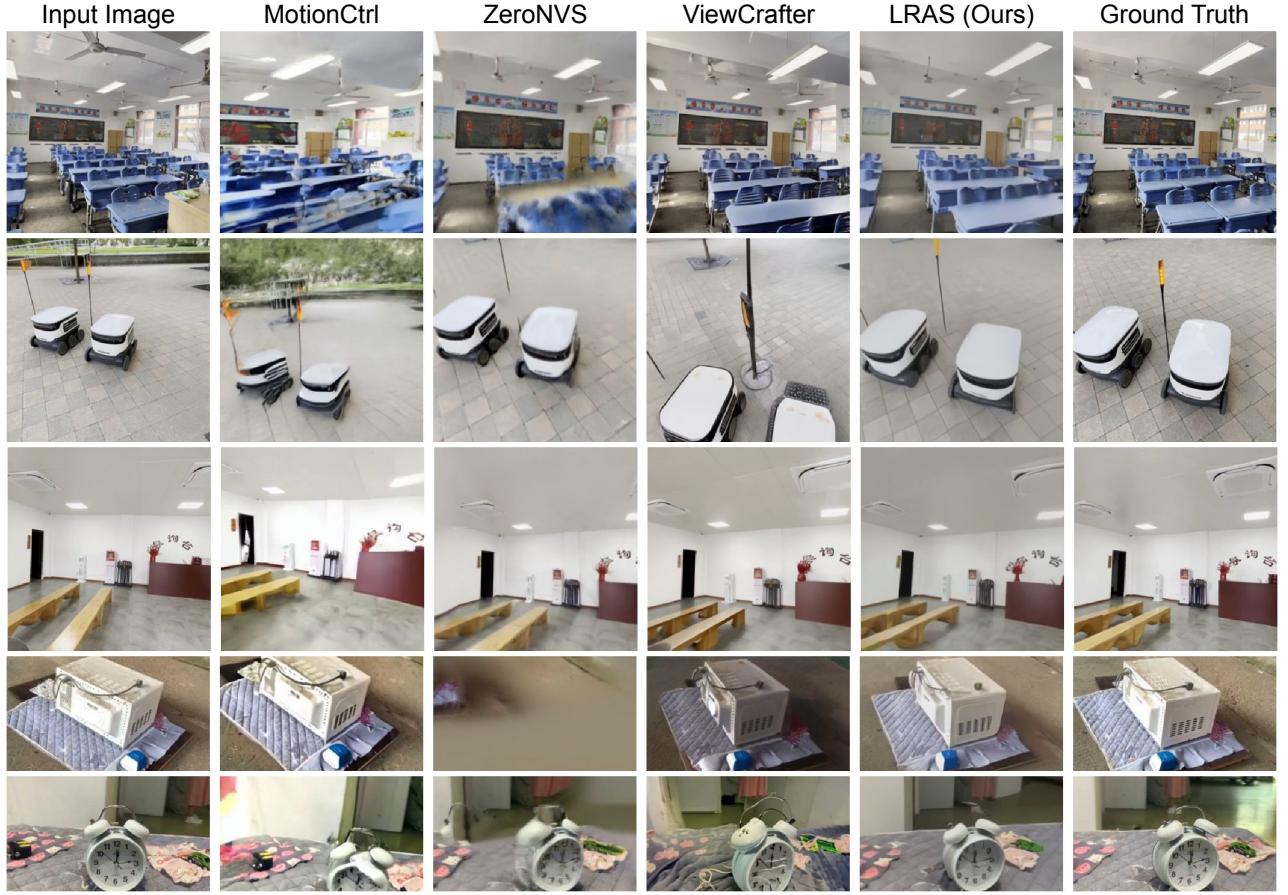


Figure 4. Novel view synthesis from a single image. The results show that our model performs controllable novel view synthesis with various camera motions in a diverse scenes. Compared to other models, the reconstructed images do not show abrupt change in object and scene identity. See supplementary for more results.

models work reasonably well for 3D manipulations.

New Object Editing Benchmark Most prior work in this area either use human evaluations on a small set of images [27], or synthetic benchmarks [31] to evaluate their method. This can be attributed to the lack of high quality real-world datasets with ground-truth 3D object transform annotations. To address this problem, we collect a dataset called **3DEditBench** consisting of 100 image pairs with a diverse set of object types undergoing rotations and translations, and inter-object occlusions. We capture these images in a variety of background and lighting conditions. To obtain the ground-truth 3D object transformation for a given pair, we annotate four corresponding points in the two images, unproject them, and use least-squares optimization to find the best-fitting rigid transformation that aligns the two sets of points. This transform is then used to create flow maps that condition **LRAS_{RGB}** to perform 3D object edits in natural scenes (see Section 4.2)

Metrics In line with our NVS evaluations in Section 4.1, we use metrics that measure the image generation quality such as PSNR, SSIM and LPIPS. However, previous

work [31] has found that these metrics often prefer image quality over edit accuracy. [31] proposed the Edit Adherance metric (EA) to directly measure of how well the boundaries of the transformed object overlaps with the ground truth. This is measured as the IOU (intersection over union) between the ground truth segment map and the estimated segment map in the generated image – we obtain these by running the SAM [19] model on these images.

Qualitative and Quantitative Comparisons We find that our model outperforms other methods on all metrics except marginally inferior LPIPS compared to Lightning-

Model	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	EA \uparrow
DragAnything	15.13	0.415	0.443	0.517
Diffusion Handles	17.82	0.567	0.344	0.619
LightningDrag	19.52	0.567	0.184	0.722
LRAS (Ours)	21.85	0.700	0.212	0.798

Table 2. Comparison of metrics for 3D object manipulation.

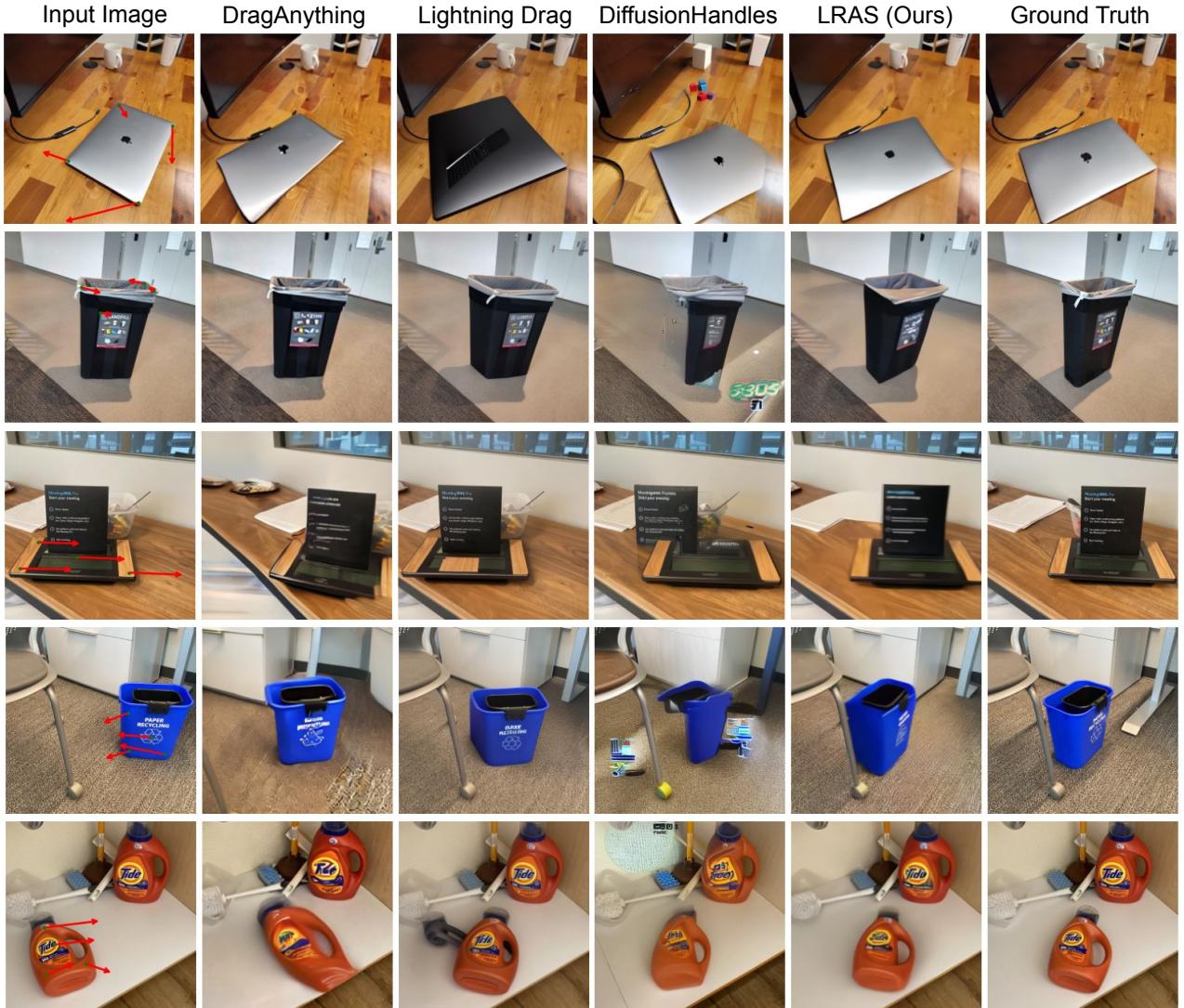


Figure 5. 3D object manipulation from a single image. We show that our model can perform both 3D object translation and rotation. Compared to the other methods, our model preserves object identity on real world images, and produces more photorealistic generated images with accurate object edits. See supplementary for more results.

Drag (see Table. 2). However, as shown in Figure. 9, we find that qualitatively our model is significantly better, especially on more complex 3D transformations. Furthermore, the Edit Adherence (EA) metric (proposed in [31]), which is a more reliable measure of the precision of the edit, seems to strongly prefer generations of our model. Interestingly, we find that DiffusionHandles [31] struggles on some of these real-world images due to failures in the null-text inversion process for natural images. The failure modes involve changing the appearance of the surrounding objects in the scene, leading to unnatural generations, blurry reconstructions, and incorrect 3D motion. A similar trend can also be seen in the drag-based image-editing baselines, albeit to a lesser degree in LightningDrag. On the other

hand, **LRAS** overcomes these limitations with autoregressive sequence modeling and generates more consistent and natural-looking images. Further, we find that our method can also be extended to perform object removal and amodal completion (we show more examples in supplementary).

4.3. End-to-End 3D Scene Understanding

4.3.1. Self-Supervised Monocular Depth Estimation

Evaluation Details We evaluate the self-supervised monocular depth estimation performance on three datasets: NYUv2 [42], BONN [30], TUM [43] datasets. NYUv2 is mostly composed of static scenes, whereas BONN and TUM include humans with implied motion. We evaluate SC-DepthV2 [2], IndoorDepth [10], and MotionCtrl [50] as

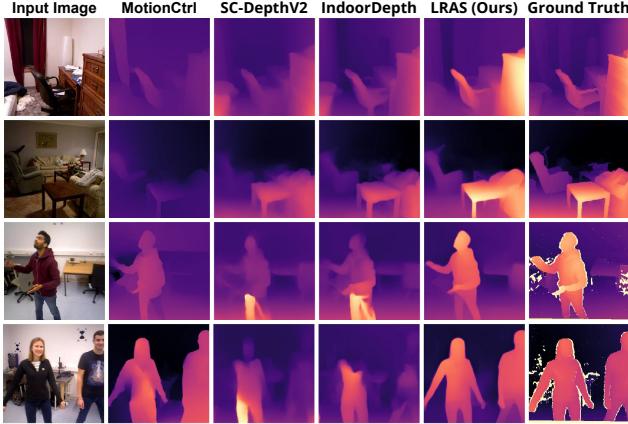


Figure 6. Self-supervised monocular depth estimation. On static scenes, our model performs comparably well to existing self-supervised depth estimation methods. However, when there are dynamic objects in the scene, our model significantly outperforms geometric-consistency-based methods, demonstrating its robustness in handling moving objects. Yellow artifacts in ground truth depth maps are noise and excluded during the evaluation.

Dataset	Model	AbsRel ↓	Log10 ↓	$\delta_1 \uparrow$
NYUD-v2	MotionCtrl	0.246	0.099	0.624
	SC-DepthV2	0.136	0.059	0.819
	IndoorDepth	0.120	0.051	0.857
	LRAS (Ours)	0.121	0.050	0.873
BONN	MotionCtrl	0.167	0.068	0.798
	SC-DepthV2	0.183	0.169	0.800
	IndoorDepth	0.167	0.064	0.827
	LRAS (Ours)	0.120	0.047	0.889
TUM	MotionCtrl	0.204	0.097	0.712
	SC-DepthV2	0.229	0.100	0.632
	IndoorDepth	0.213	0.094	0.682
	LRAS (Ours)	0.179	0.073	0.766

Table 3. Comparison of metrics for self-supervised monocular depth estimation.

baselines. To extract depth from MotionCtrl, we induce an upward in-plane camera motion and compute the disparity between the first and 7th images using RAFT [47]. As our model uses square images as input, we center-crop the images for all datasets and only evaluate the square region.

Qualitative and Quantitative Comparisons Our results demonstrate that **LRAS** achieves high-quality depth reconstruction in both static and dynamic settings, as shown in Figure 6. The baseline models exhibit limitations because they rely on static geometry consistency, preventing them from extracting training signals from moving objects. In contrast, our model successfully learns depth cues from optical flow. While optical flow is not always purely induced

by camera motion, we empirically found that averaging optical flow while moving the camera upward leads to reliable depth estimation. MotionCtrl demonstrates better generalization to dynamic objects than other self-supervised methods, but lacks strong depth understanding overall, as evident by its weaker performance in static scenes. Table 3 confirms our observations quantitatively, where our model achieves competitive performance on NYUv2, and outperforms other methods on dynamic datasets, BONN and TUM. Overall, our findings in self-supervised depth estimation highlight the importance of optical flow as a prediction target. The results also strengthen the argument for autoregressive modeling, where simple modification of sequence design can naturally facilitate other tasks.

4.3.2. End-to-end Novel View Synthesis

Dataset	LRAS	PSNR ↑	SSIM ↑	LPIPS ↓
WildRGB-D	w. Our Depth	16.716	0.484	0.264
	w. DA-V2	17.748	0.536	0.218
DL3DV	w. Our Depth	17.984	0.516	0.332
	w. DA-V2	18.110	0.523	0.328

Table 4. Comparison of metrics for novel view synthesis depending on depth model.

Since our framework can estimate depth, we explored using our model instead of a supervised depth model to create a fully self-supervised NVS pipeline. Table 4 presents the results of NVS using depth predicted by our model. While the metrics generally declined compared to the pipeline using Depth Anything V2 (DA-V2), the drop was not severe. This indicates that our model’s depth estimation is sufficiently accurate for novel view synthesis, reinforcing the feasibility of a unified, self-supervised 3D vision framework with optical flow and autoregressive training.

5. Discussion & Conclusion

In this work, we introduce **LRAS**, an autoregressive sequence modeling framework with local patch quantization and random access prediction. We show that our method outperforms diffusion-based models in 3D editing capabilities, ensuring consistency in objects and scenes during editing. The model also offers precise camera control and object manipulation, demonstrating a strong understanding of spatial relationships and transformations in 3D. Furthermore, we demonstrate that our modeling framework is flexible. With a simple change in sequence design, it can leverage optical flow either as input conditioning for 3D scene editing or as a prediction target for depth estimation.

Overall, **LRAS** provides a robust and scalable alternative to diffusion models for 3D scene understanding, expanding the potential of autoregressive modeling in vision. Future

work could explore the integration of additional modalities to further enhance spatial and physical reasoning.

6. Acknowledgment

This work was supported by the following awards: To D.L.K.Y.: Simons Foundation grant 543061, National Science Foundation CAREER grant 1844724, National Science Foundation Grant NCS-FR 2123963, Office of Naval Research grant S5122, ONR MURI 00010802, ONR MURI S5847, and ONR MURI 1141386 - 493027. We also thank the Stanford HAI, Stanford Data Sciences and the Marlowe team, and the Google TPU Research Cloud team for computing support.

References

- [1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Hermann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3
- [2] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. Auto-rectify network for unsupervised indoor depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):9802–9813, 2021. 7
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- [4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2
- [5] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19457–19467, 2024. 2
- [6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 2
- [7] Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey et.al. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024. 12
- [8] Patrick Esser, Robin Rombach, and Björn Ommer. Tampering transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2020. 3
- [9] NVIDIA et. al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 3
- [10] Chao Fan, Zhenyu Yin, Yue Li, and Feiqing Zhang. Deeper into self-supervised monocular indoor depth estimation. *arXiv preprint arXiv:2312.01283*, 2023. 7
- [11] Wanquan Feng, Tianhao Qi, Jiawei Liu, Mingzhen Sun, Pengqi Tu, Tianxiang Ma, Fei Dai, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol: Disentangled and unified video motion synthesis control. *arXiv preprint arXiv:2411.17765*, 2024. 3
- [12] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 3
- [13] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 4
- [14] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *ArXiv*, abs/2501.03847, 2025. 4
- [15] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025. 3
- [16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 3
- [17] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and SungHyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. *arXiv preprint arXiv:2502.08244*, 2025. 3
- [18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4, 12
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5, 6
- [20] Juil Koo, Paul Guerrero, Chun-Hao Paul Huang, Duygu Ceylan, and Minhyuk Sung. Videohandles: Editing 3d object compositions in videos using video generative priors. *arXiv preprint arXiv:2503.01107*, 2025. 2
- [21] Mathis Koroglu, Hugo Caselles-Dupré, Guillaume Jeanneret Sanmiguel, and Matthieu Cord. Onlyflow: Optical flow based motion conditioning for video diffusion models. *arXiv preprint arXiv:2411.10501*, 2024. 3
- [22] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few

- images using transformers. In *European Conference on Computer Vision*, pages 198–216. Springer, 2022. 2
- [23] Tianhong Li, Qinyi Sun, Lijie Fan, and Kaiming He. Fractal generative models. 2025. 4
- [24] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2025. 2, 4
- [25] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 4
- [26] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2
- [27] Oscar Michel, Anand Bhattacharjee, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems*, 36:3497–3516, 2023. 6
- [28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 12
- [29] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 3
- [30] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019. 7
- [31] Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J Mitra. Diffusion handles enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7695–7704, 2024. 2, 4, 5, 6, 7
- [32] Arnaud Pannatier, Evann Courdier, and François Fleuret. σ -gpts: A new approach to autoregressive models, 2024. 4
- [33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 13
- [34] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 12
- [36] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 4
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [38] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022. 2
- [39] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. *arXiv preprint arXiv:2310.17994*, 2023. 2, 12
- [40] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Y. Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Da, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *ArXiv*, abs/2401.15977, 2024. 4
- [41] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent YF Tan, and Jiashi Feng. Lightningdrag: Lightning fast and accurate drag-based image editing emerging from videos. *arXiv preprint arXiv:2405.13722*, 2024. 2, 5
- [42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 7
- [43] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgbd slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 7
- [44] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 46(1):497–508, 2023. 2
- [45] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2
- [46] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristof

- fersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 12
- [47] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4, 8, 12
- [48] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2, 12
- [49] Zhouxia Wang, Yushi Lan, Shangchen Zhou, and Chen Change Loy. Objctrl-2.5 d: Training-free object control with camera poses. *arXiv preprint arXiv:2412.07721*, 2024. 2
- [50] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 7
- [51] Daniel Winter, Matan Cohen, Shlomi Fruchter, Yael Pritch, Alex Rav-Acha, and Yedid Hoshen. Objectdrop: Bootstrapping counterfactuals for photorealistic object removal and insertion. In *European Conference on Computer Vision*, pages 112–129. Springer, 2024. 13
- [52] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024. 2, 5
- [53] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 5
- [54] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 4, 12
- [55] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2
- [56] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2
- [57] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 2
- [58] Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David C. Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. 2023. 3
- [59] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2
- [60] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 4
- [61] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. *arXiv preprint arXiv:2412.01821*, 2024. 3
- [62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5
- [63] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 2
- [64] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 4

3D Scene Understanding Through Local Random Access Sequence Modeling

Supplementary Material

A. Dataset

We collect a large dataset of diverse video clips crawled from the internet, totaling about 7,000 hours in length, called big video dataset.

The videos were crawled using LLaMA 3 [7]-generated search queries about videos that contain lots of physical dynamics, diverse settings and objects. Specifically, the crawl search queries were generated using Kinetics400 [18] action categories, and supplemented with additional sport and physical activity categories, as well as product review categories. The videos were filtered to contain some minimal amount of optical flow, and to align with predefined CLIP [35] keyword filters. Our positive CLIP keywords are “action,” “activity,” “motion,” and “place,” and our negative keywords are “animation,” “cartoon,” “face,” “game menu,” “graphic,” “map,” “newscast,” “person,” and “screenshot.” Alignment is measured as the dot product between the CLIP embeddings of keywords and video frames.

To improve camera motion diversity in the **LRAS_{FLOW}** training data, we converted 280 scenes from ScanNet++ [54] into Neural Radiance Fields [28, 46] and rendered videos from them with known diverse camera trajectories. Discretized relative camera pose change between two frames is provided to **LRAS_{FLOW}** as conditioning when available in the data.

B. NVS Evaluation Details

To evaluate novel view synthesis, we compare generated images to ground-truth real-world images using known camera poses. While camera rotation is unambiguous, camera translation may have arbitrary scale. Therefore, it is necessary to find the right scene scale to perform fair evaluations for all of the models.

To align MotionCtrl and ZeroNVS results with ground-truth images, we sweep a range of scene scales and take the generated trajectories with the best median LPIPS score across frames. For ZeroNVS, we sweep scales in the range 0.1 to 10, multiplying the scale by the ground-truth camera translations from each evaluation dataset. ZeroNVS introduces a normalization scheme [39] at training time to address this scale ambiguity, but does not apply it at inference. For MotionCtrl, we sweep the range 1 to 10, as smaller translation scales empirically weaken the camera conditioning and lead to incorrect camera pose trajectories. Scale alignment for these models may fail for samples with especially poor 3D reconstruction quality. For ViewCrafter, we resolve the scene scale using their method of aligning point clouds with DUST3R [48]. For **LRAS**, we have computed

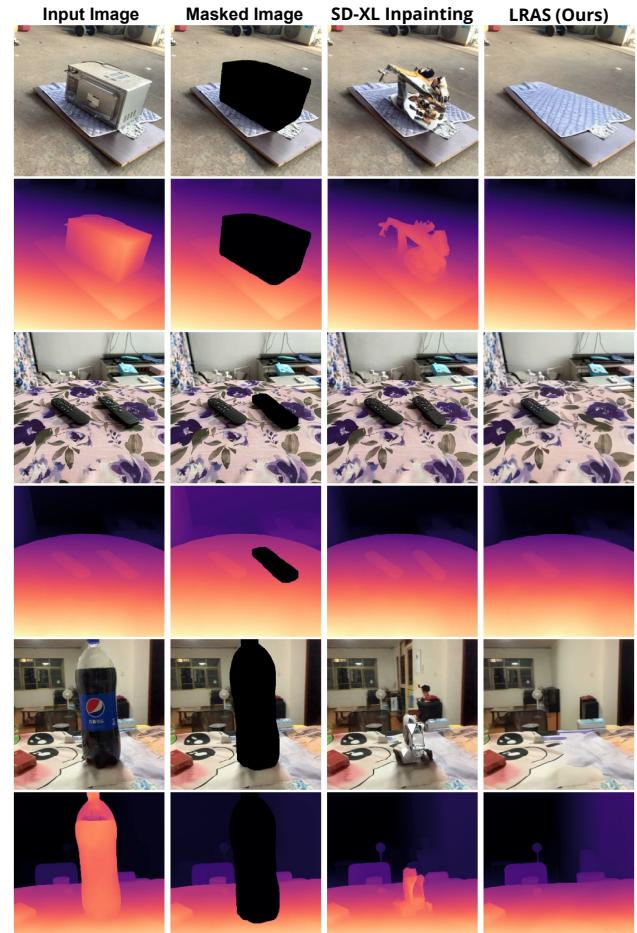


Figure 7. Amodal Completion. We compare our self-supervised amodal depth reasoning with the inpainting method. The inpainting method struggles with underdetermined scene changes, as it lacks explicit control over object removal. In contrast, the flow-based physical scene editing approach conditions object removal more precisely, resulting in more reliable amodal reasoning.

the single scale value per scene by matching the optical flow computed from the video using RAFT [47] and the 2D flow computed from the depth and relative camera pose changes.

Since ViewCrafter operates on wide rectangular videos, we adapt the input images accordingly. For DL3DV, which consists of wide images, we provide the full image to ViewCrafter. For WildRGB-D, which contains narrower images, we provide a center-cropped rectangular region to ViewCrafter. All other models receive a center-cropped square image as input for both datasets. All evaluation metrics are computed only on the overlapping regions; for

WildRGB-D, this region is rectangular, and for DL3DV it is square.

C. Amodal Completion

A simple yet powerful application of our model is amodal reasoning. By applying high-magnitude flow to the object, we effectively remove it from the scene. We compare this approach to a self-supervised heuristic for object removal based on image inpainting using the Stable Diffusion XL (SD-XL) [33] model. As shown in Fig. 7, our model successfully removes objects while reconstructing the occluded regions with reasonable accuracy. In contrast, the SD-XL approach may struggle with imperfect segmentation or implicit object presence caused by shadows or nearby objects. This problem is also observed by other work [51], where they address it by introducing a specific dataset for training. Our method, however, provides an explicit physical cue for object removal via optical flow, enabling more controlled and interpretable amodal reasoning in a self-supervised way.

D. Additional Qualitative results on NVS and object manipulations

In Figure 8 we include additional qualitative results for novel view synthesis and in Figure 9 we include additional qualitative results for object manipulation.

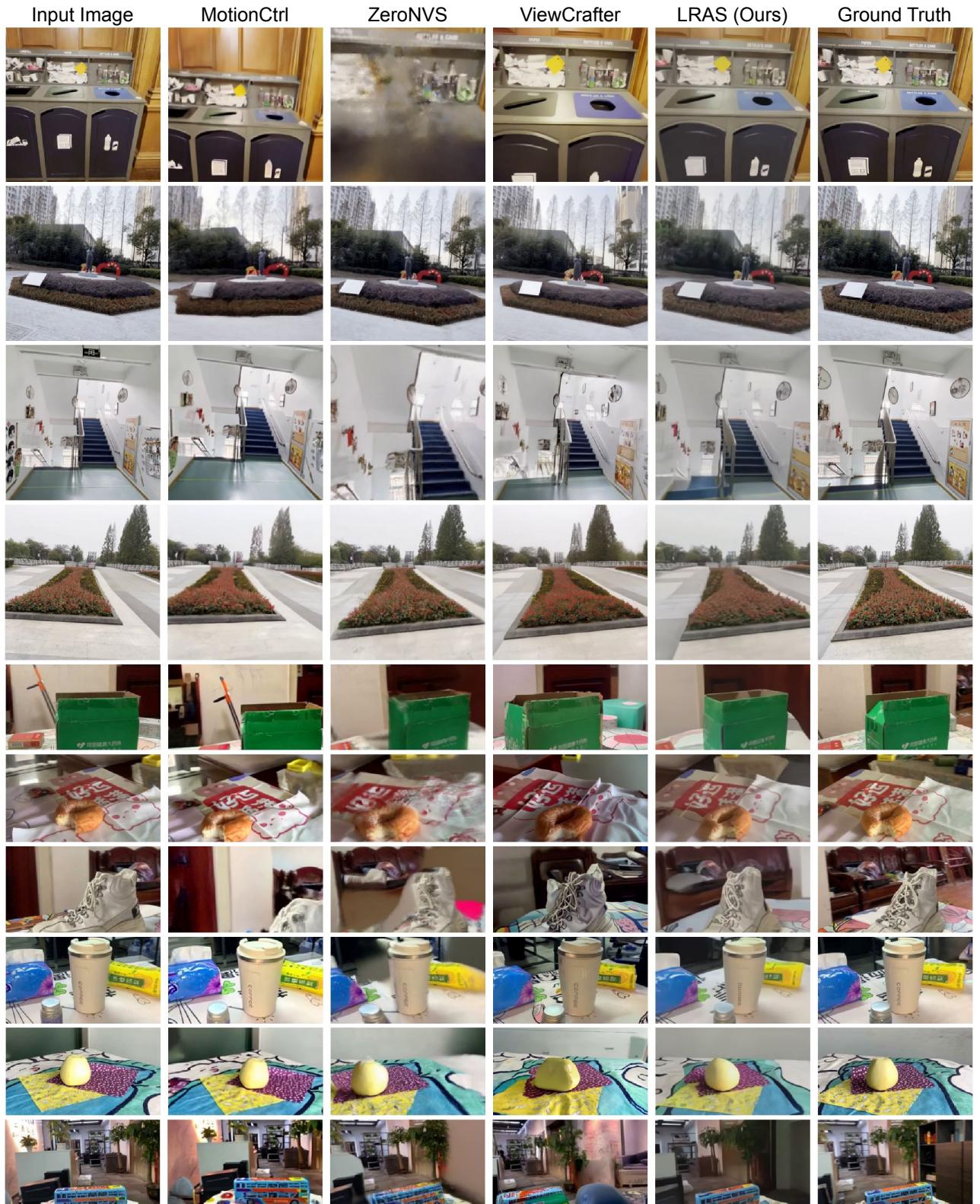


Figure 8. Additional results on novel view synthesis from a single image. The results show that our model performs controllable novel view synthesis with various camera motions in a diverse scenes. Compared to other models, the reconstructed images do not show abrupt change in object and scene identity.

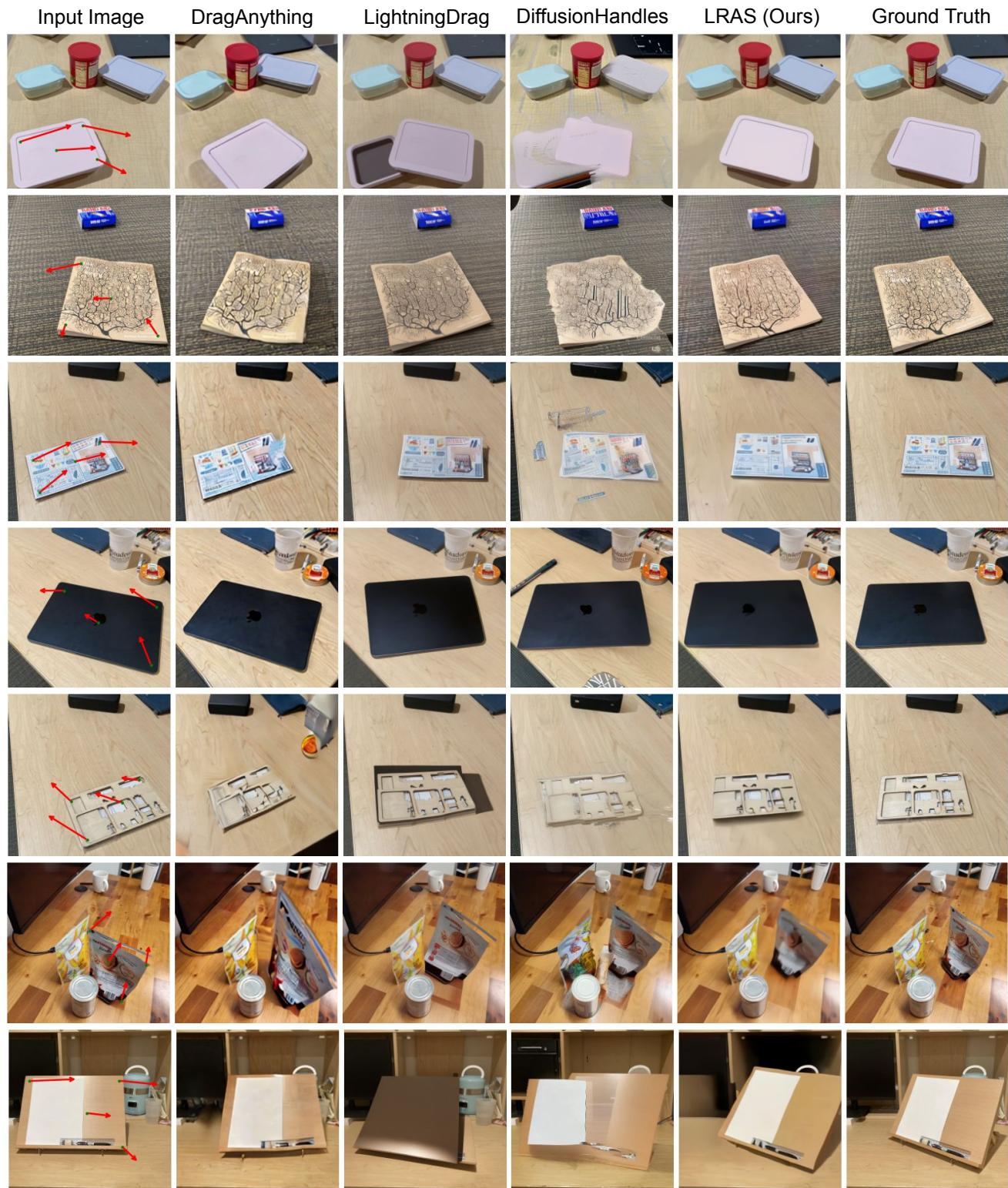


Figure 9. Additional results on 3D object manipulation from a single image. The results show that our model can perform both object translation and rotation in 3D. Compared to the other methods, our model does not change the object identity even for in-the-wild real world images.