ALIBABA DAMO ACADEMY

Project Page

NEURAL INFORMATION PROCESSING SYSTEMS

MVInpainter: Learning Multi-View Consistent Inpainting to Bridge 2D and 3D Editing

Chenjie Cao[1,2,3], Chaohui Yu[2,3], Fan Wang[2,3], Xiangyang Xue[1], Yanwei Fu[1*]

[1]Fudan University, [2]DAMO Academy, Alibaba Group, [3]Hupan Lab

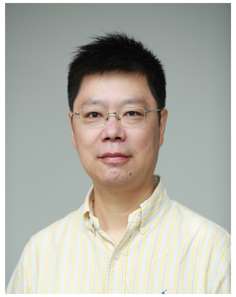
{caochenjie.ccj,huakun.ych,fan.w}@alibaba-inc.com

{xyxue,yanweifu}@fudan.edu.cn

Chenjie Cao

Chaohui Yu

Fan Wang

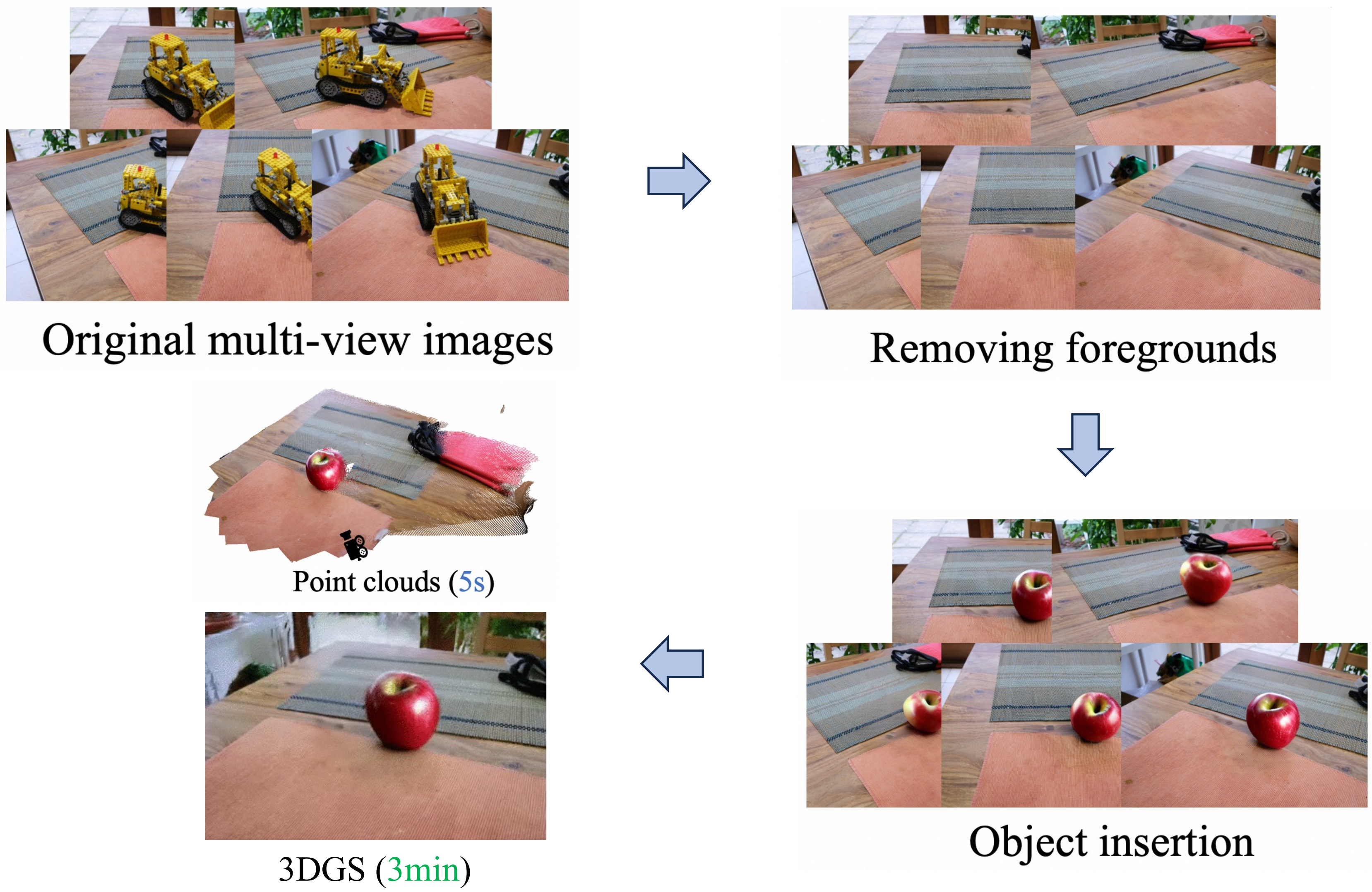


Xiangyang Xue

Yanwei Fu

Our Focus: Real-world 3D editing

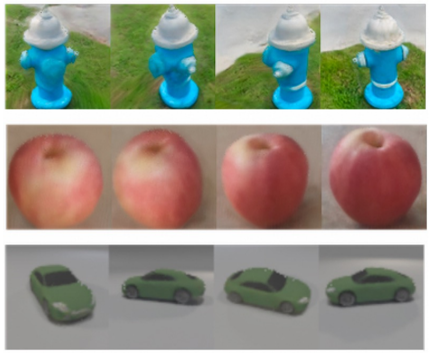
Removal, insertion, and replacement



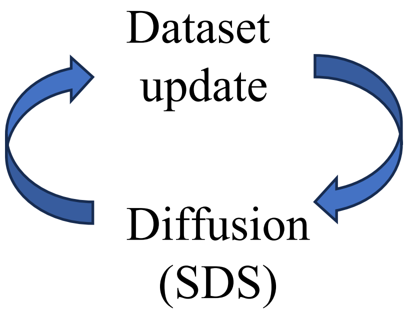
Challenges

- 3D object generation struggles to generalize to scene-level editing
- Novel view synthesis methods have difficulty generalizing across various categories
- Instance-level 3D editing is time-consuming
- Heavy reliance on explicit camera poses

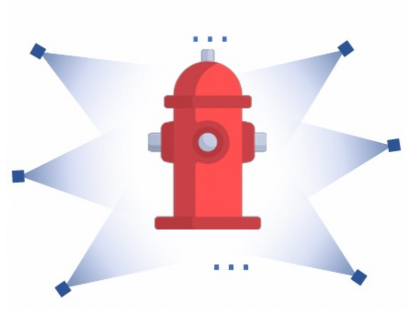
Focusing object-centric synthetic data



Limited scene categories



Time-consuming

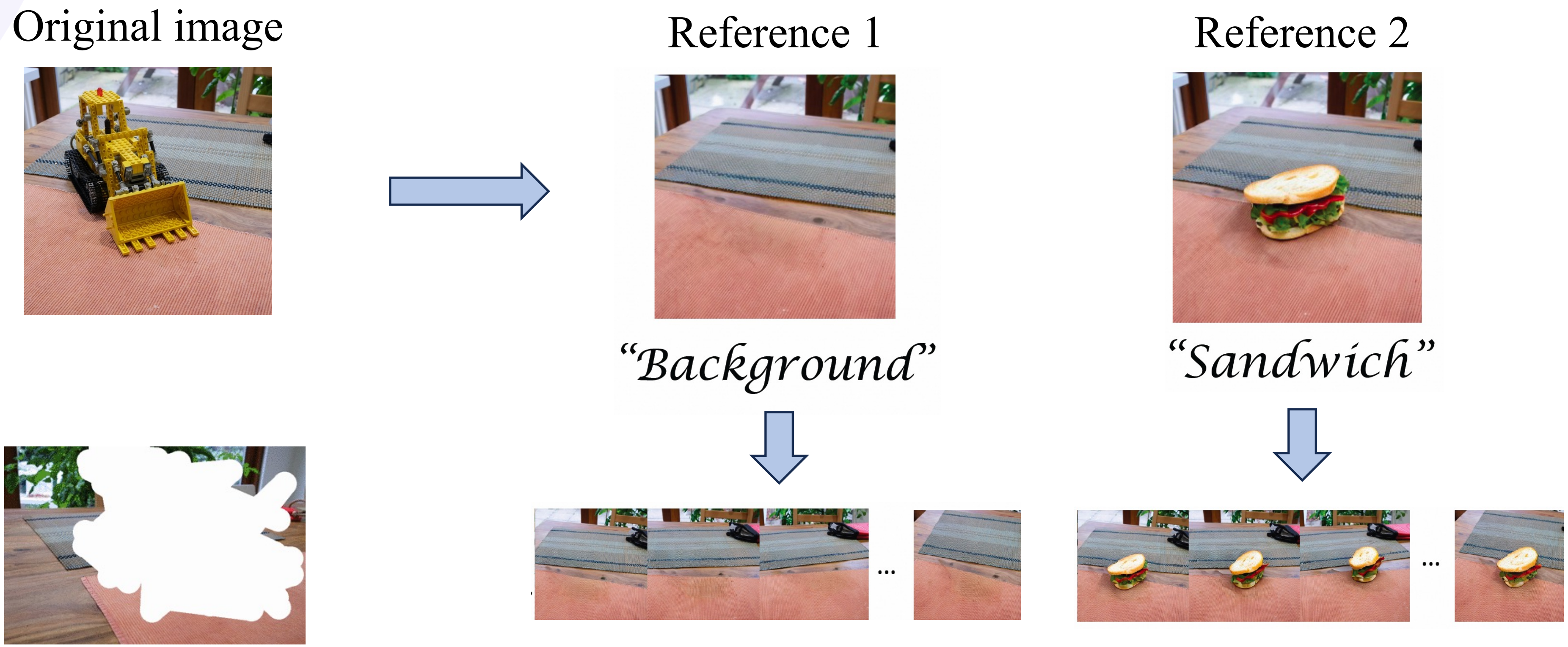


Explicit camera requirements

Key motivation

達摩院 DAMO ACADEMY

2D-inpainting enjoys good performance with large text-to-image models

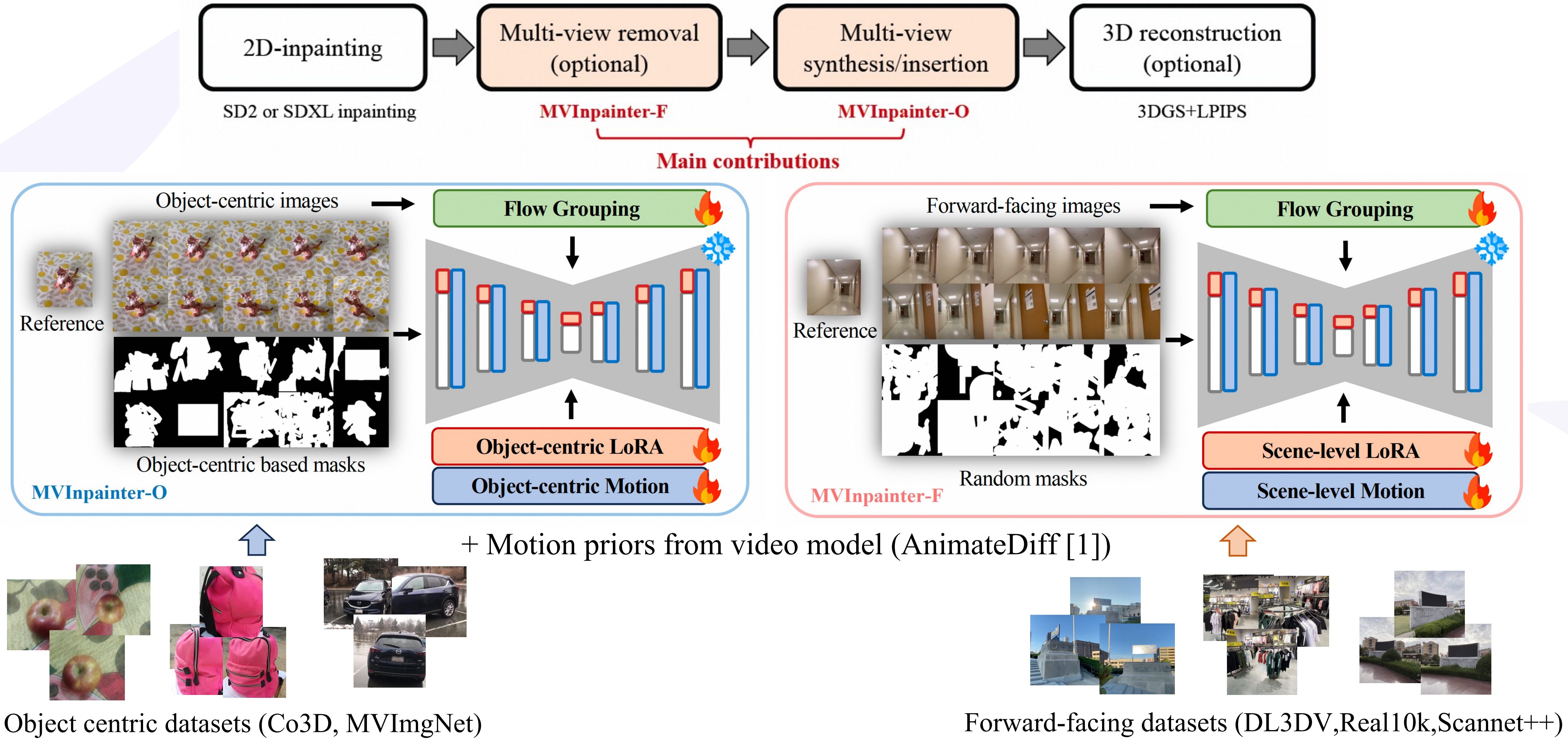


MVInpainter enables 3D editing with a **multi-view consistent inpainting** manner, effectively extending 2D generation into 3D scenarios.

Consistency and identity preserving

The overall framework of MVInpainter

达摩院 DAMO ACADEMY

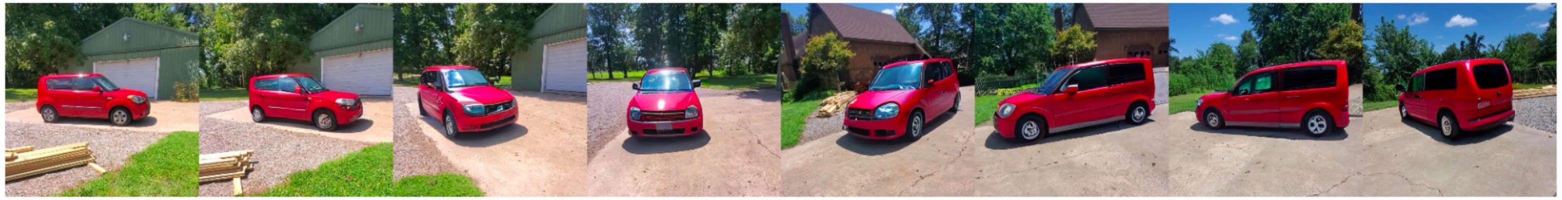


[1] Guo Y, Yang C, Rao A, et al. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. ICLR, 2024.

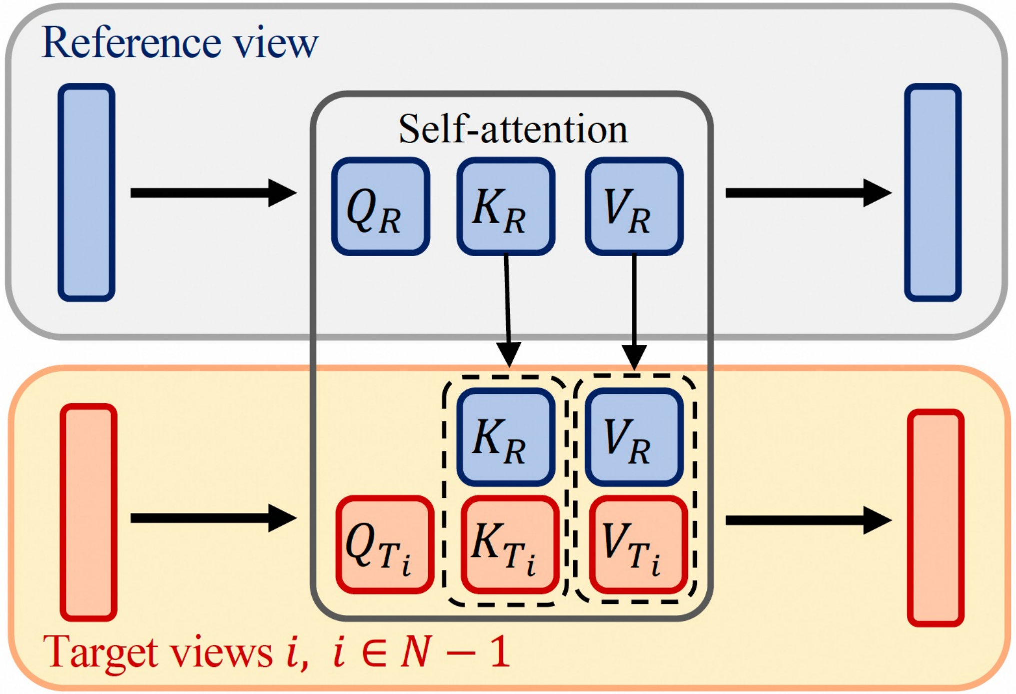
Motion Priors from Video Models


Reference and masked inputs

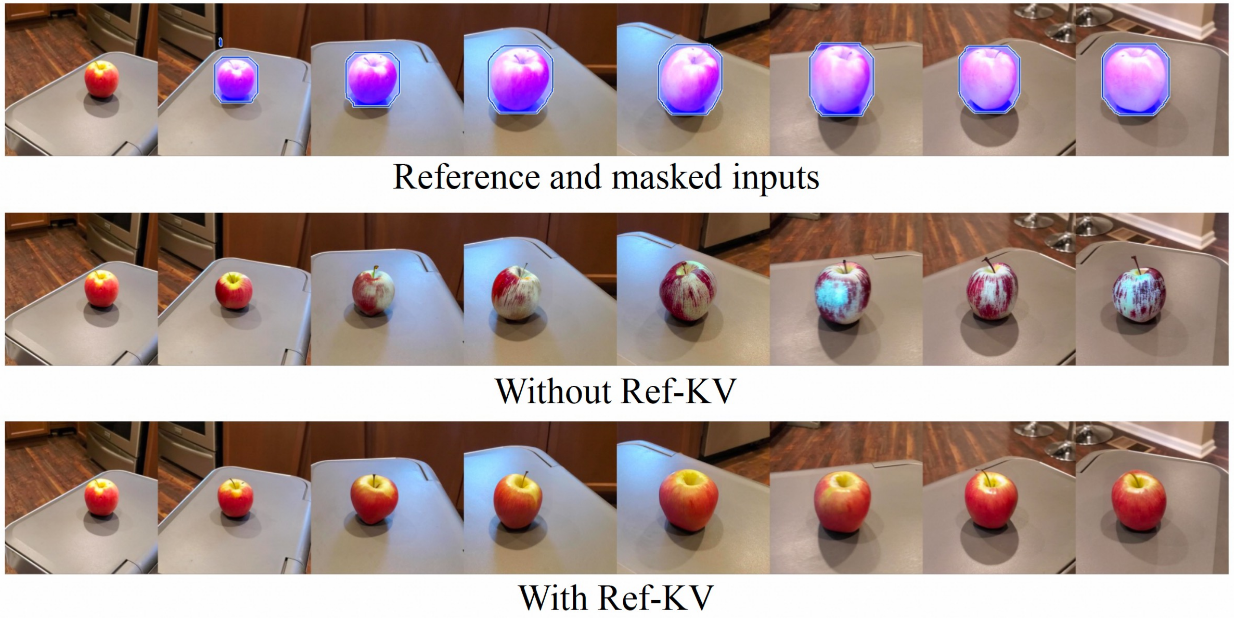

Without video priors (AnimateDiff)


With video priors (AnimateDiff)

Reference Key&Value Concatenation (Ref-KV)

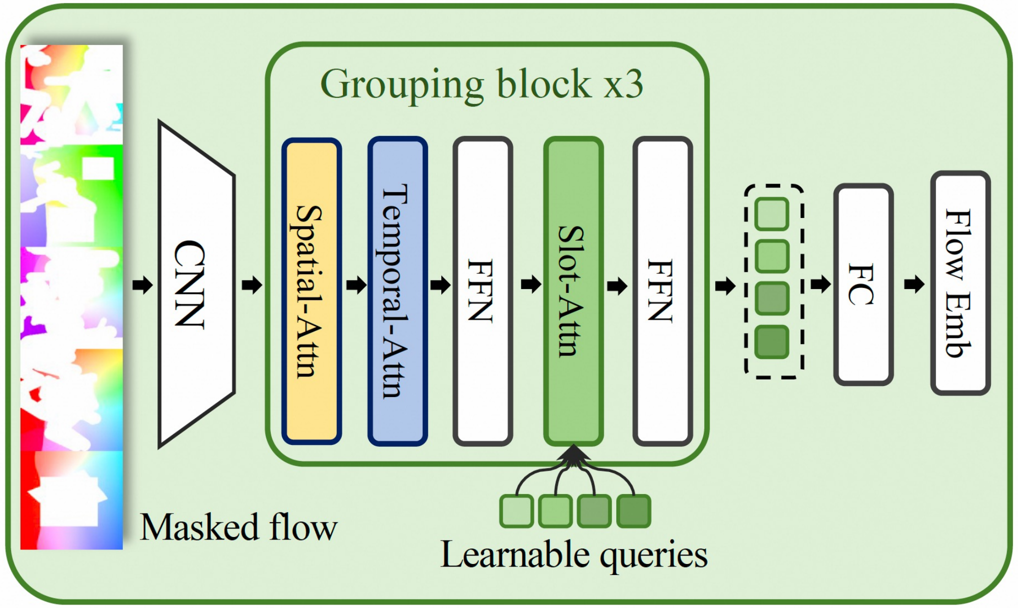


(b) Ref-KV of the self-attention block in U-Net

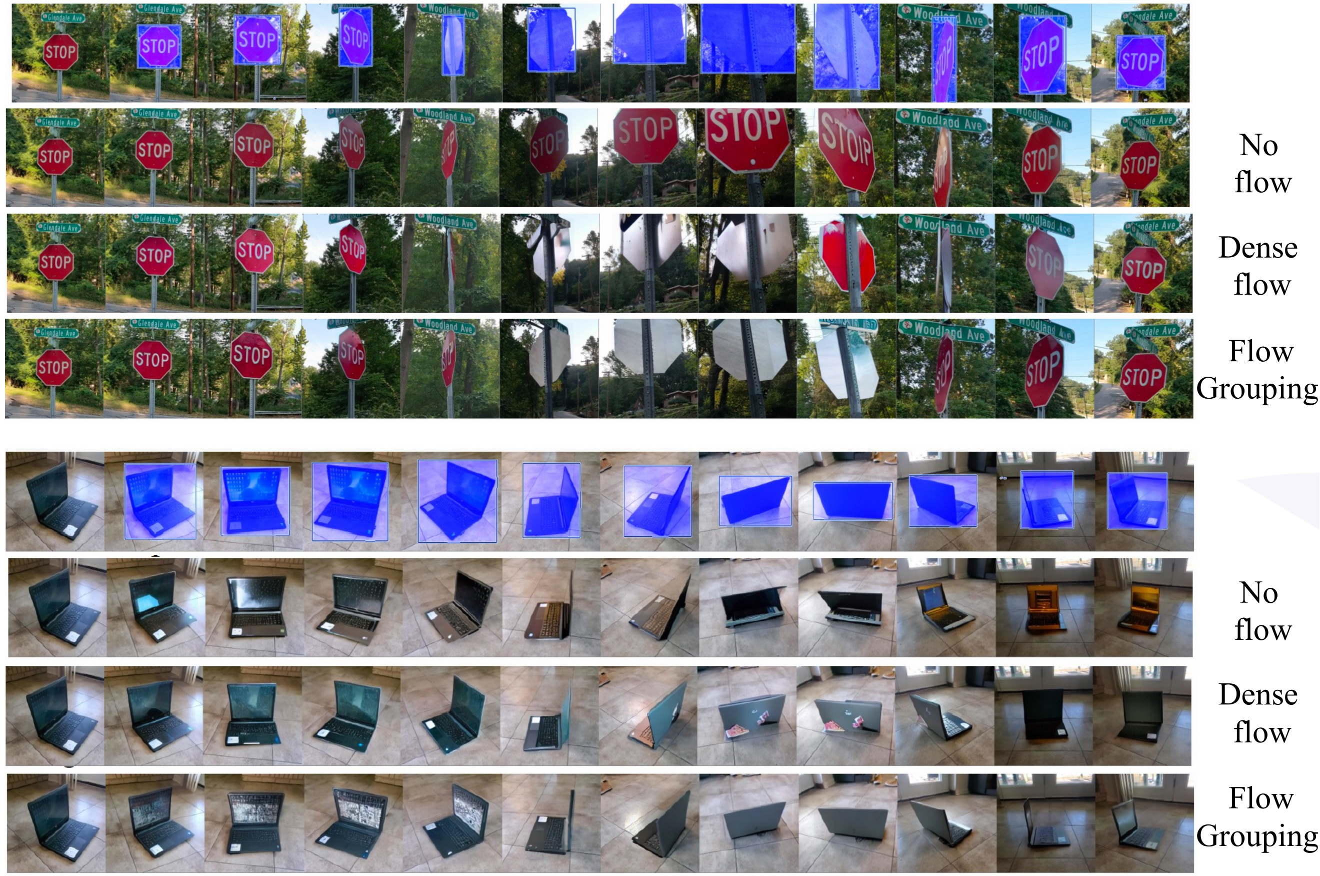


Pose-Free Flow Grouping

We utilize the slot-attention to learn high-level flow features for implicit camera control



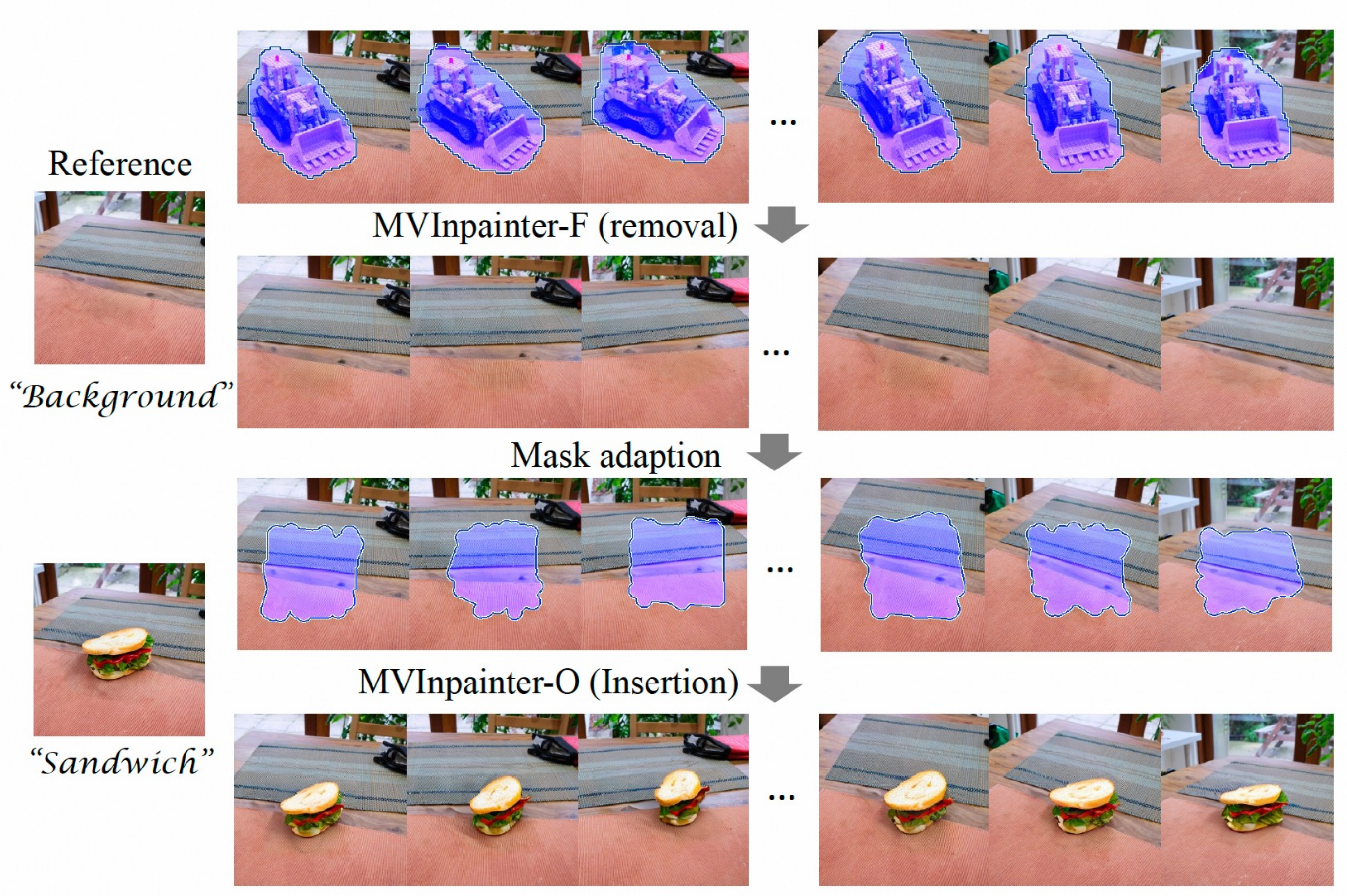
(c) Architecture of Flow Grouping



Flow grouping outperforms dense flow! (avoiding overfitting inaccurate flow estimation)

Inference Pipline

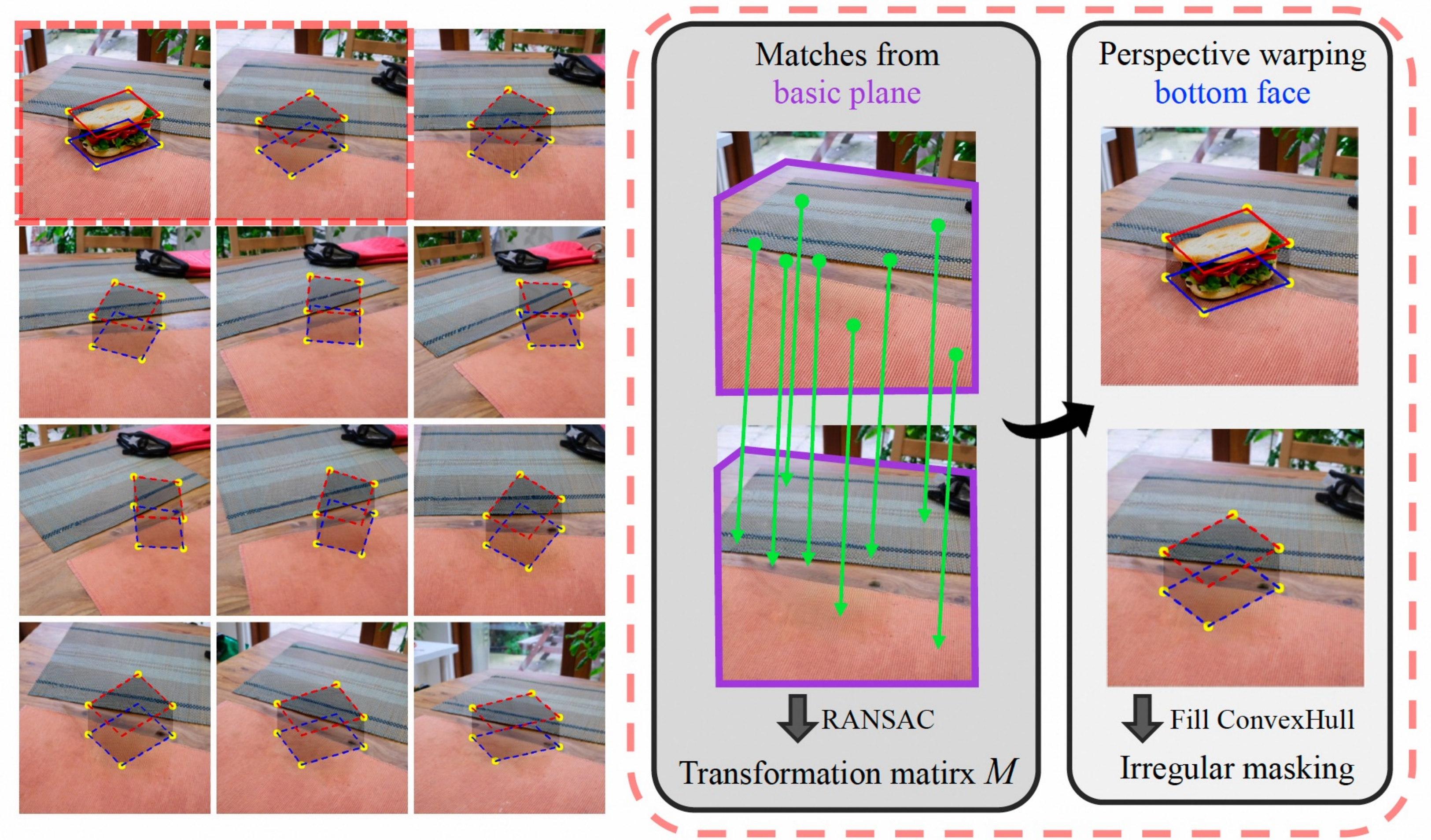
達摩院 DAMO ACADEMY



(a) Inference pipeline

How to achieve mask locations in inference?

Mask Adaption



(b) Masking adaption (c) Perspective warping of the object plane

Assumption: the 3D box's bottom face and the basic plane on which the object is placed must be the same plane

So these two planes share the same perspective transformation

Implementation Details

- Training setup: 8 A800 GPUs; batch size 64; learning rate 1e-4; MVInpainter-O 100k steps; MVInpainter-F 60k steps; dynamic frame fine-tuneing 10k steps

- Metrics: PSNR, SSIM, LPIPS, CLIP-score, FID, KID, and DINOv2 similarity (DINO-A, DINO-M)

- Training frame numbers: frame number 12; dynamic frame number 8~24

- Datasets:
  - Object-centric: Co3D, MVImgNet
  - Forward-facing: DL3DV, Real10k, Scannet++

Experiments

達摩院
DAMO ACADEMY

Object-centric NVS results

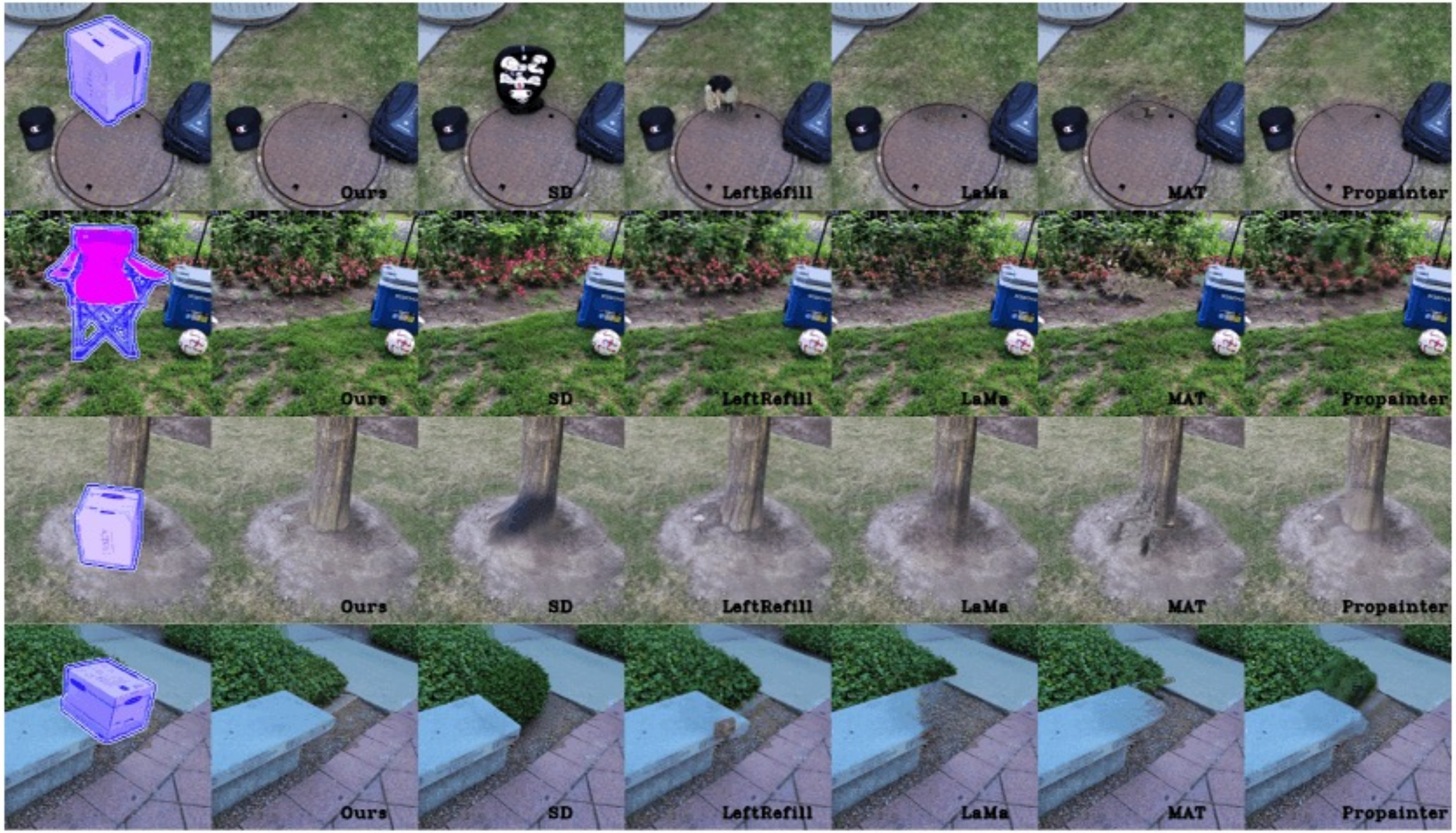
	CO3D+MVImgNet					Omni3D (zero-shot)				
ZeroNVS [64]	12.44	0.606	41.90	0.981	0.6028	9.38	0.627	82.81	5.421	0.5451
Nerfiller [80]	18.29	0.310	36.64	0.491	0.6603	16.10	0.272	37.04	1.056	0.6279
LeftRefill [7]	17.74	0.283	38.06	0.826	0.6392	17.09	0.239	27.81	0.775	0.6484
Ours	**20.25**	**0.185**	**17.56**	**0.154**	**0.8182**	**19.19**	**0.153**	**16.40**	**0.345**	**0.7667**

Forward-facing inpainting results

	SPInNeRF (removal)					Scannet+Real10K+DL3DV (inpainting)			
	PSNR↑	LPIPS↓	FID↓	DINO-A↑	DINO-M↑	PSNR↑	LPIPS↓	FID↓	KID↓
LaMa [70]	28.62	0.054	15.26	0.8909	**0.6019**	17.61	0.337	38.47	0.981
MAT [37]	27.05	0.067	28.81	0.8727	0.5760	15.47	0.377	37.38	0.899
SD-inpaint [59]	26.98	0.070	19.32	0.8556	0.4422	13.54	0.417	38.67	1.048
LeftRefill [7]	30.29	0.102	18.02	0.8931	0.5652	15.14	0.380	38.06	1.334
ProPainter [102]	**31.72**	0.047	12.25	0.8757	0.5534	20.42	0.306	61.76	2.642
Ours	28.87	**0.036**	**7.66**	**0.8972**	0.5937	**20.91**	**0.173**	**15.58**	**0.252**

Object-centric NVS

Object removal



Compared to NeRF Editing

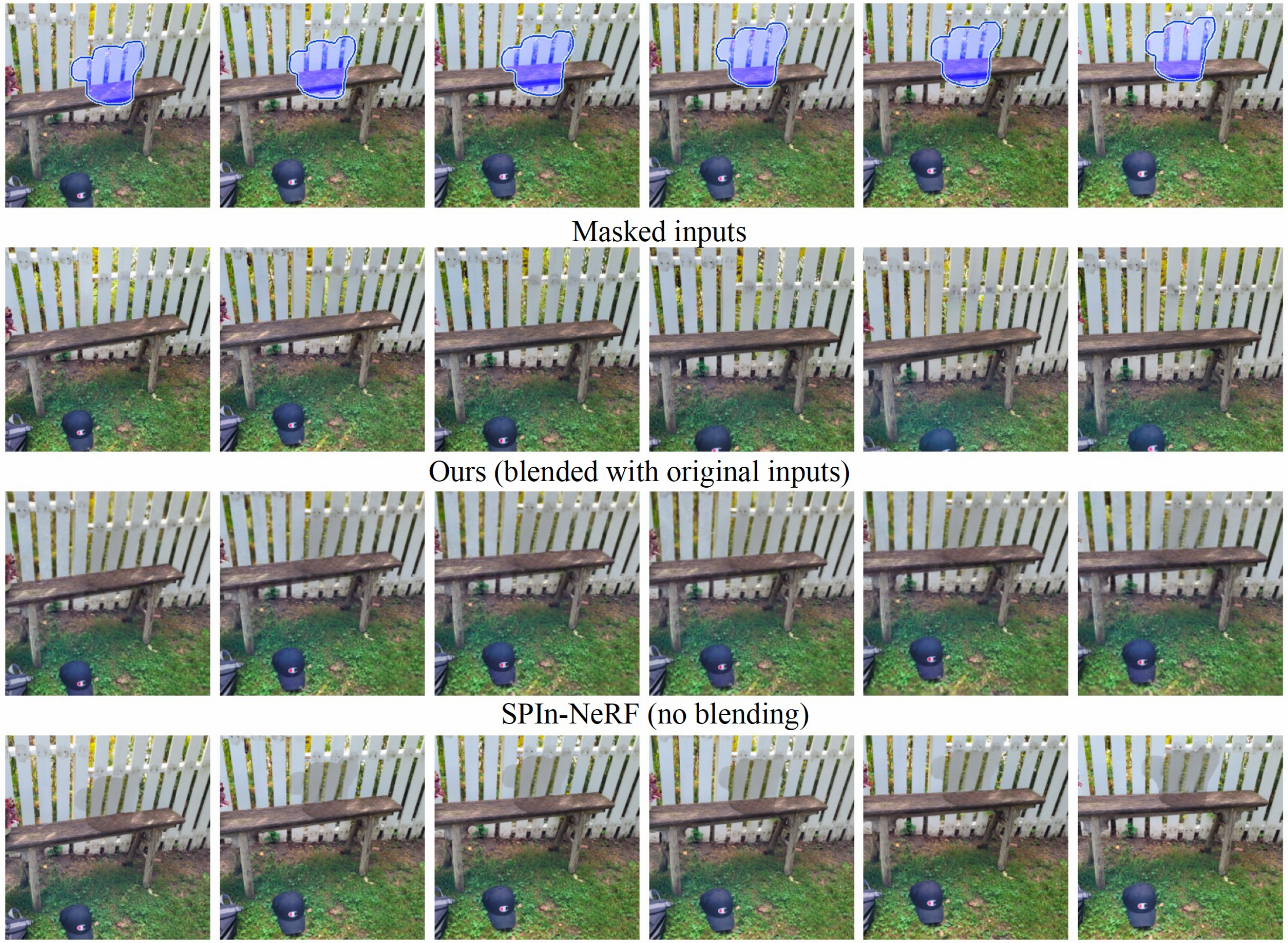
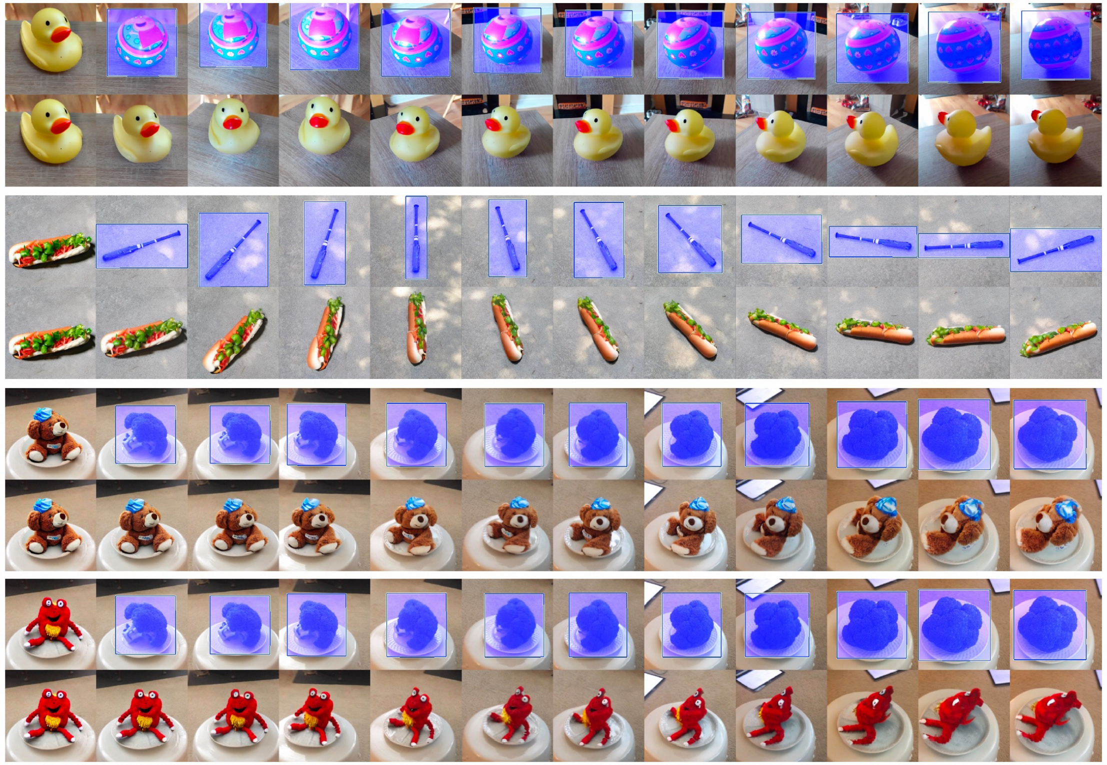


Figure 19: Object removal compared to SPIn-NeRF [51].

	PSNR↑	LPIPS↓	FID↓	DINO-S↑	DINO-L↑
Ours	**28.87**	**0.036**	**7.66**	**0.8972**	0.5937
SPIn-NeRF	25.82	0.084	38.13	0.8681	**0.6350**

Object replacement

AnyDoor

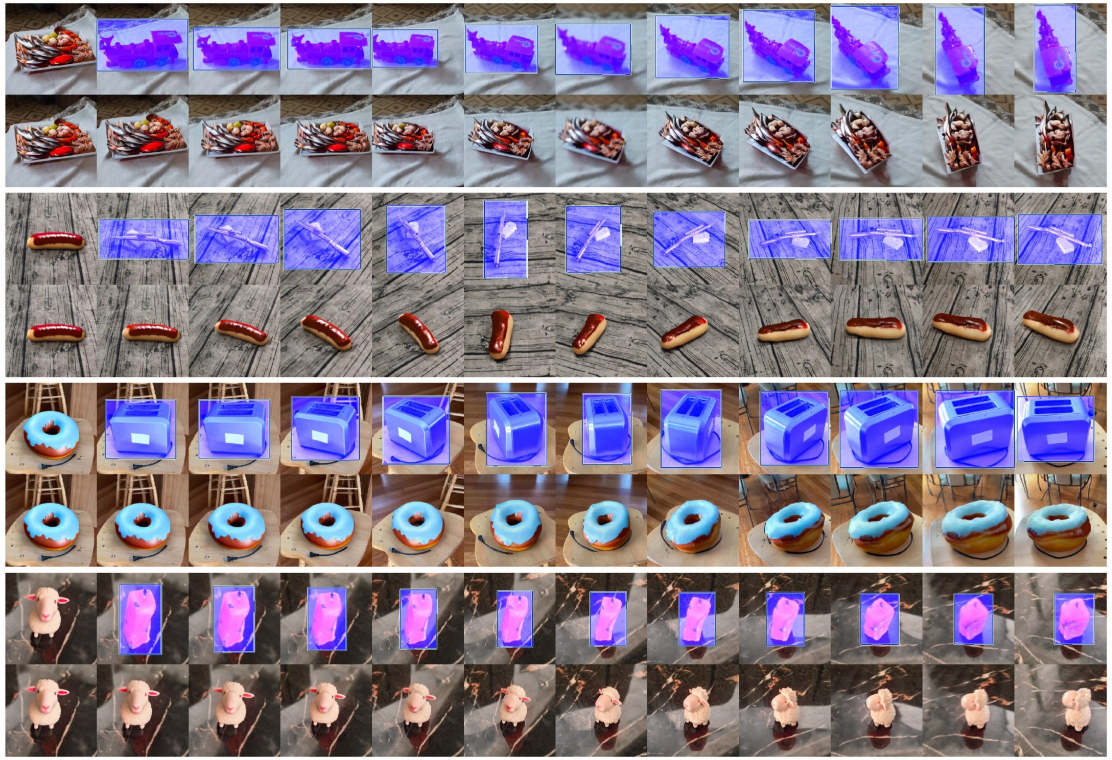


“seafood”

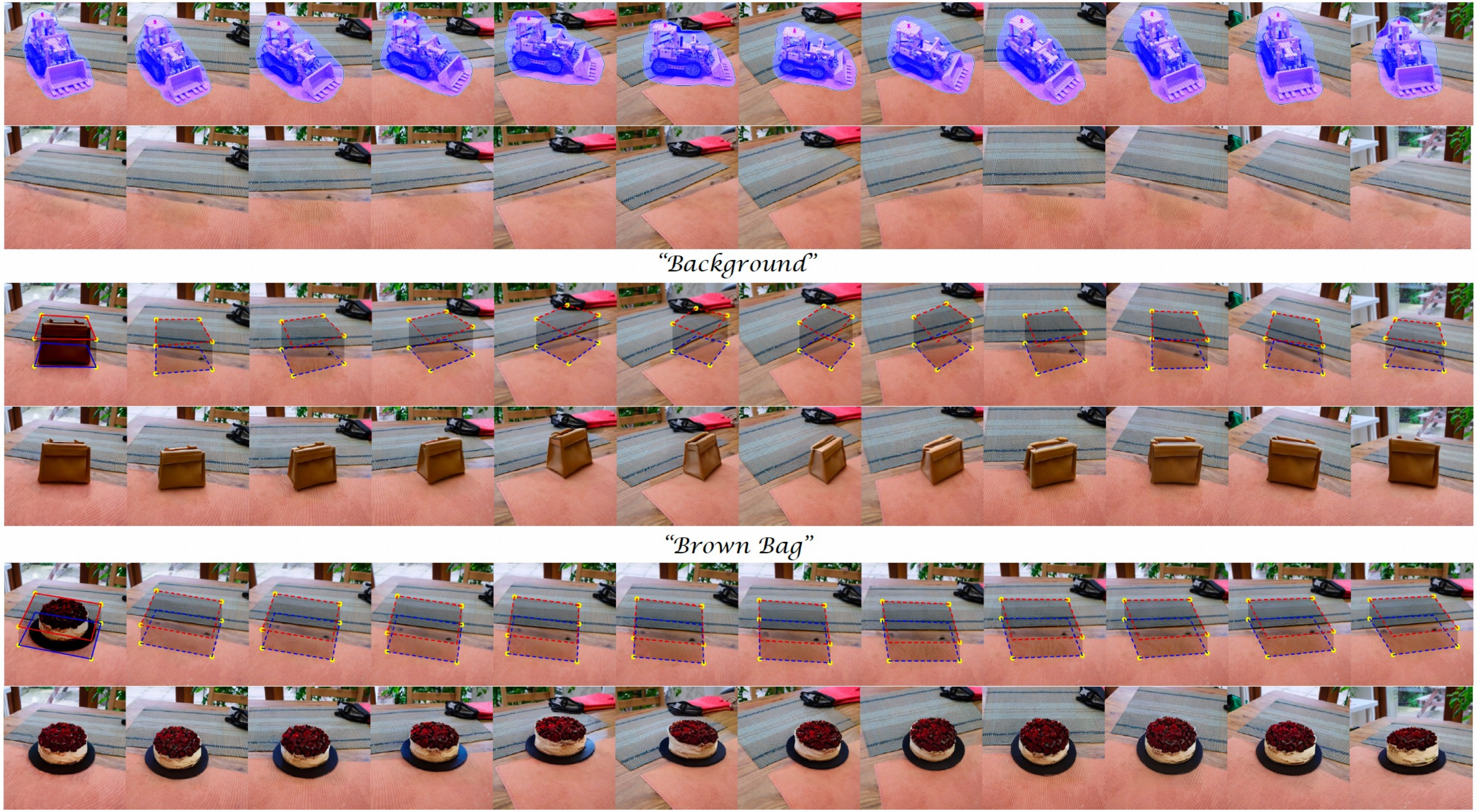
“bread”

“Donut”

“Sheep toy”



Scene editing



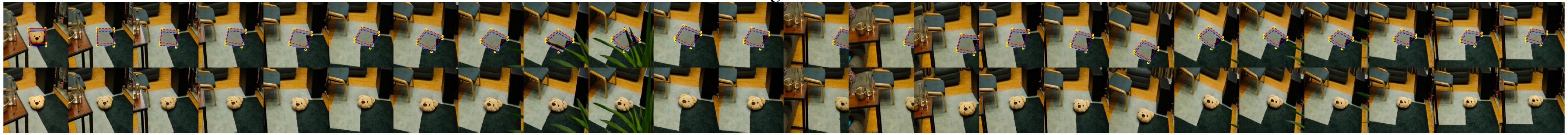
"Background"

"Brown Bag"

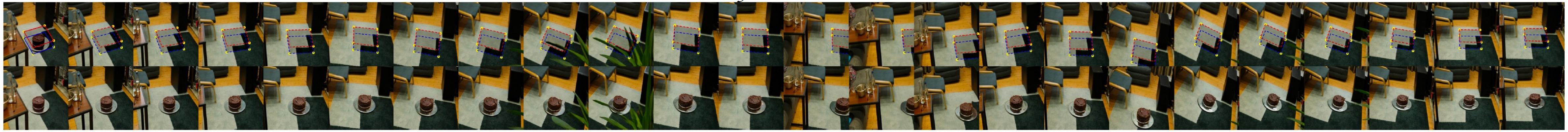
"Chocolate Cake"

Scene editing

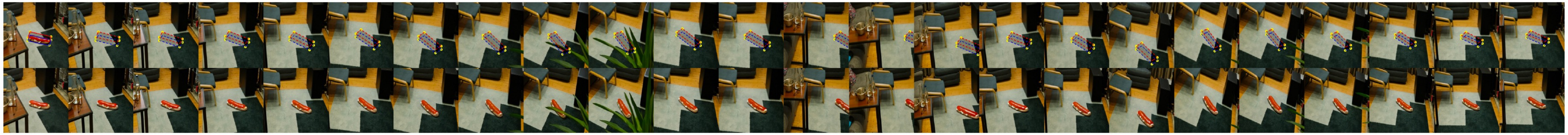
“Background”



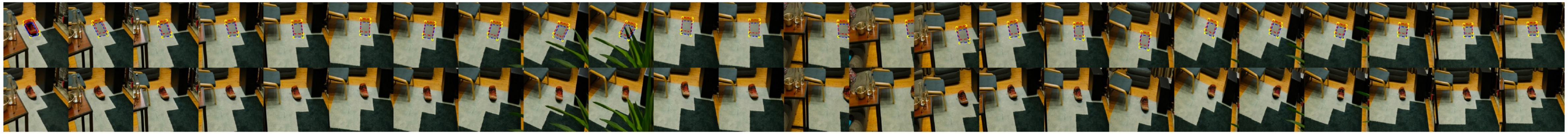
“Teddy Bear’s Head”



“Chocolate Cake”



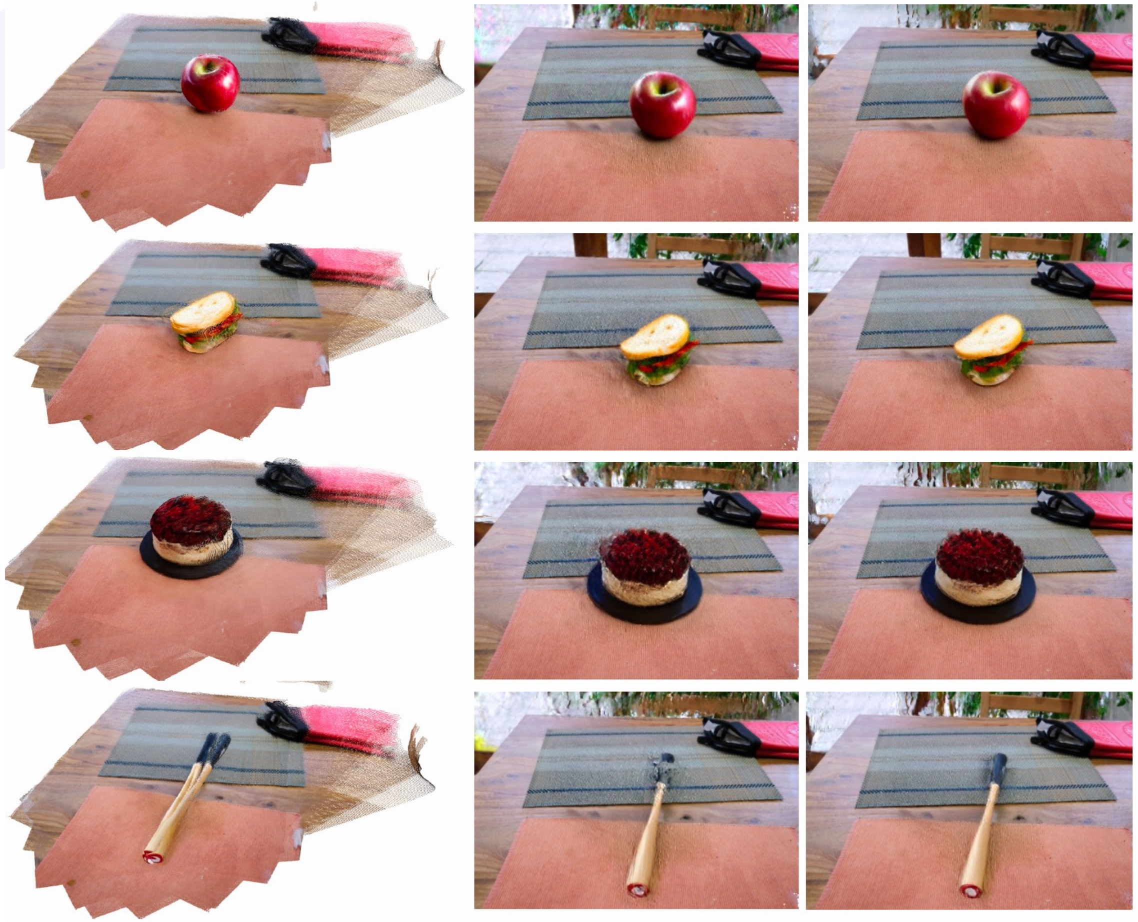
“Hotdog with Tomato Paste”



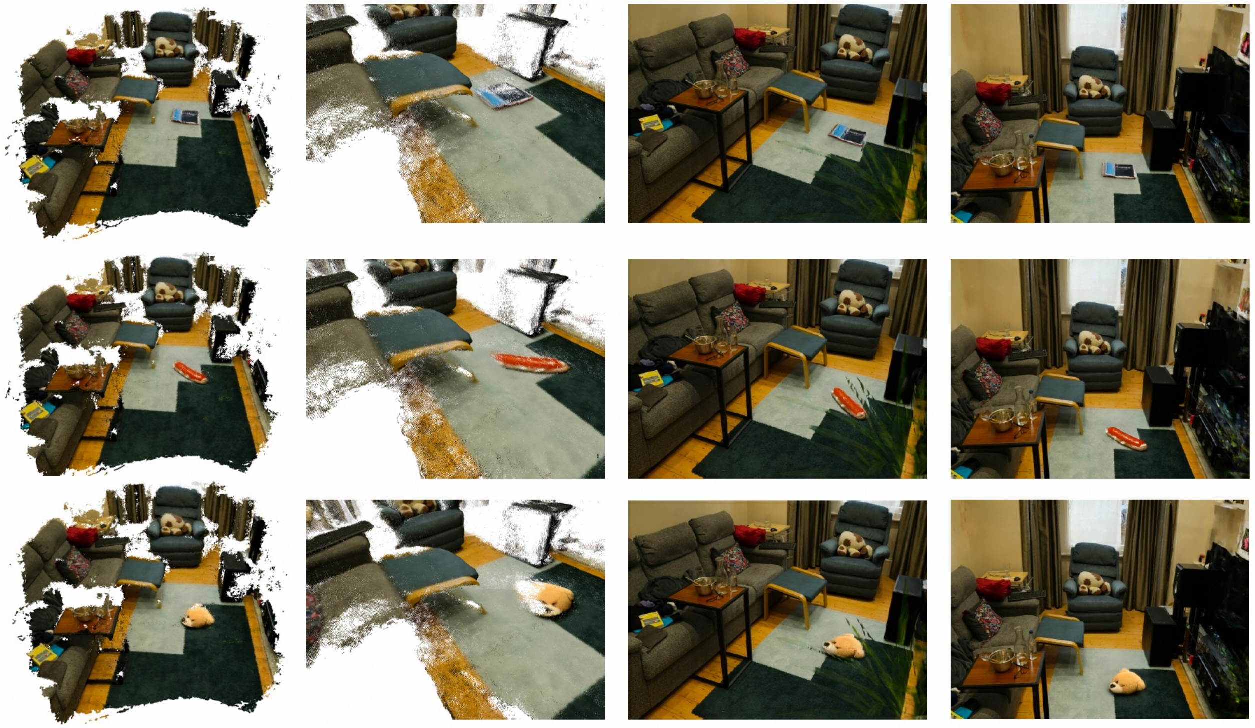
“Orange Shoe”

3DGS reconstruction

We initialize the point cloud through Dust3R or MVS.
The 3DGS is optimized by L1, SSIM, and masked LPIPS losses



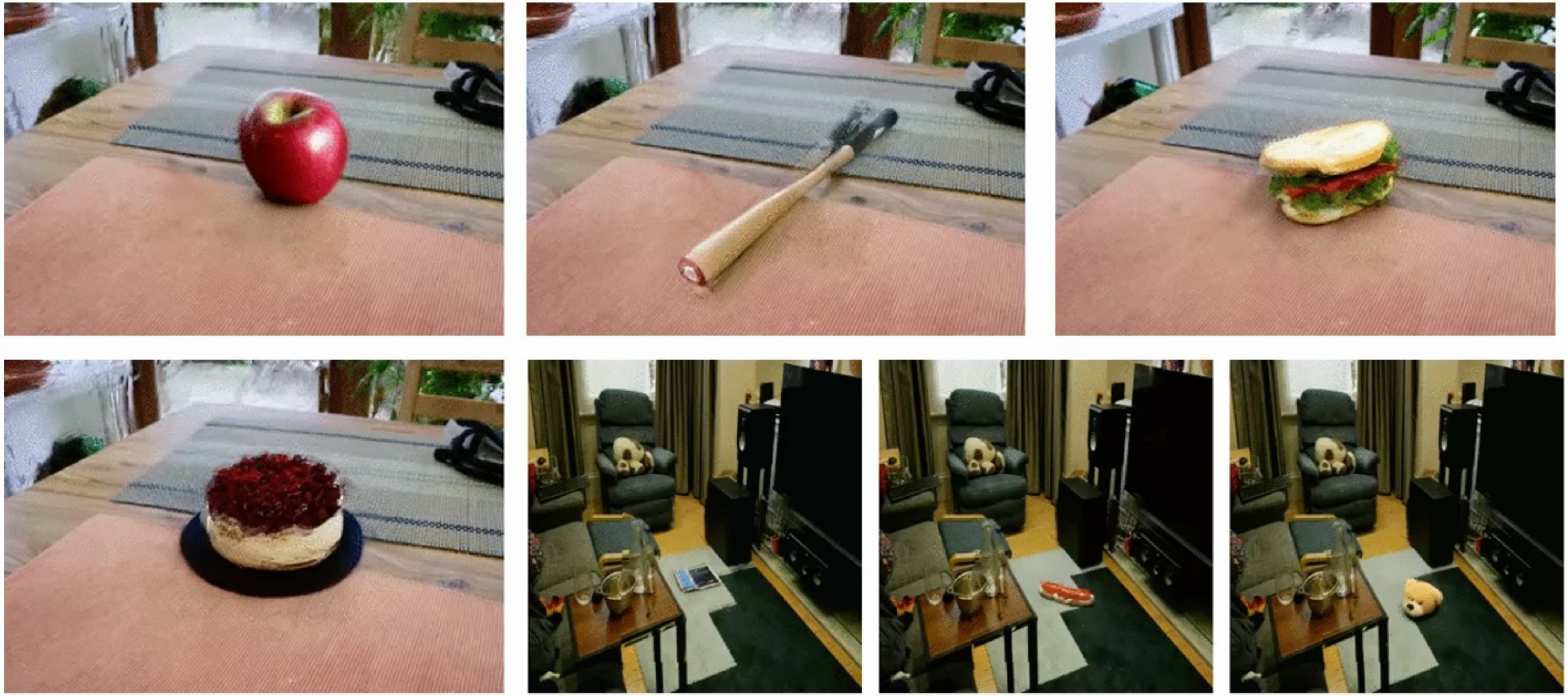
(a) Point clouds (b) Test view w/o LPIPS (c) Test view with LPIPS



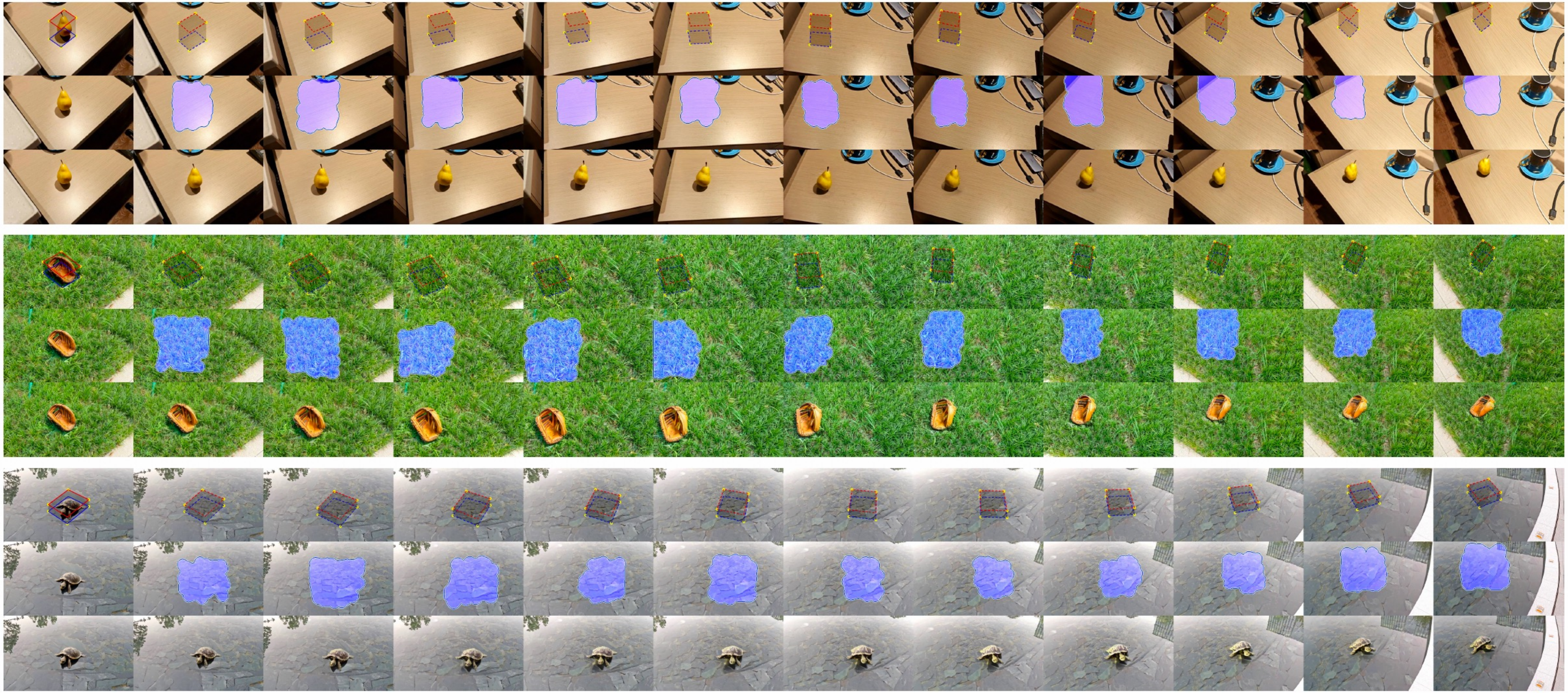
(a) MVS point clouds (b) 3DGS test views

NVS results

3DGS results



Robustness of mask adaption



Ablation study and Efficiency

Table 3: Ablation studies on CO3D. 'w.o. inp' means the baseline without the inpainting formulation.

	PSNR↑	LPIPS↓	CLIP↑
Baseline	17.16	0.305	0.750
Baseline (w.o. inp)	14.35	0.443	0.648
+AnimateDiff	17.31	0.308	0.756
+Ref-KV	17.90	0.283	0.773
+Object mask	18.64	0.250	0.796
+Flow emb	**18.93**	**0.240**	**0.798**

(a) Ablation results of different proposed components

	PSNR↑	LPIPS↓	CLIP↑
No Flow	18.64	0.250	0.796
Dense Flow	18.53	0.247	0.792
Slot2D Flow (time-emb)	18.74	0.244	**0.798**
Slot2D Flow (cross-attn)	18.81	0.245	0.796
Slot3D Flow (cross-attn)	**18.93**	**0.240**	**0.798**

(b) Ablation of various strategies to inject flow guidance

Table 5: Ablation study of the baseline method with inpainting formulation, and without inpainting formulation (SD-blend and SD-NVS).

	PSNR↑	LPIPS↓	CLIP↑
SD-blend	14.35	0.443	0.648
SD-NVS	11.61	0.663	0.677
Baseline	**17.16**	**0.305**	**0.750**

Table 8: Inference time cost tested on A800 NVIDIA GPU. The view number is 24, while all inputs are resized into 256×256.

Methods	Ours	AnimateDiff	Nerfiller	LeftRefill
DDIM steps	50	50	20	50
Time	11.5s	10.1s	32.4s	33.0s

Summary

- MVInpainter is a multi-view consistent inpainting method to expand 2D generations into 3D scenes by multi-view object removal, insertion, and replacement.

- Motion initialization based on video priors and Ref-KV are presented to facilitate the structure and appearance consistency respectively.

- MVInpainter is camera-free. The flow grouping based on the slot-attention is used to encourage implicit motion control.