PoseTraj: Pose-Aware Trajectory Control in Video Diffusion

Supplementary Material

A. Synthetic Data Consturction

A.1. Construction Pipeline

In this section, we describe the detailed pipeline for constructing synthetic data. First, we set up a realistic virtual scene featuring a fixed camera, a wood-textured floor, and indoor HDRI images to simulate natural indoor environments, including floor texture and lighting. Next, we sample an object from a filtered subset of 2,000 objects from Objaverse [1]. The object is then normalized to a height of 1 unit and placed on the floor.

We then generate a random trajectory by defining a curve with a randomly initialized starting point, rotational angle, and length. The starting points of the trajectories are randomly sampled within a circle of radius 1 unit, centered at the origin $(0, 0)$. The initial orientation of the object is set at a random angle between 0° and 90° relative to the positive x-axis. Two types of trajectory templates are defined: i) a circular trajectory without any turning points, and ii) an 'S'-shaped trajectory with one turning point. For both trajectory types, the radius of rotation is uniformly sampled between 1 and 1.5 units, and the corresponding rotation angle is set between 90° and 180°. The object is animated to follow this trajectory over 200 steps while maintaining a fixed rotation center to simulate rotational motion. The movements are rendered at 5 fps with 32 keyframes using Blender's Cycles engine, with each object sampled between $1 \sim 8$ times. Fig. A2 presents visualizations of the animated data with various rotational trajectories and objects.

A.2. Ablation Study on Data scale

To ensure sufficient diversity and robustness for pretraining, we carefully evaluated our dataset of 2,000 objects with sampled videos, determining that this amount approaches the model's capacity limit, as shown in Tab. A1 and Fig. A1.

Table A1. Comparison results on the synthetic validation set of our designed ablation studies on synthetic data scale during the pretraining stage.

Vids	Objs	Synthetic-Val		
		ObjMC↓	FID ↓	FVD↓
1,000	200	0.1987	48.05	190.35
2,000	400	0.2065	47.19	187.12
5,000	1,000	0.1960	46.62	185.47
10,000	2,000	0.1895	46.34	186.01

Ablation Study on syntehtic data scale. To ensure effective pretraining as a pose-aware 3D understanding injection, it is critical to collect a sufficient amount of data for robust model learning. To investigate the optimal data size for two-stage pose-aware pretraining, we conducted an ablation study examining the impact of the number of training videos and the corresponding rendered objects. Specifically, we extended the number of videos from 1,000 to 10,000, scaling the quantity of rendered objects accordingly. As presented in Tab. A1, the model demonstrates a reliable rotational understanding with 5,000 or more training videos, whereas it struggles to learn diverse rotational patterns with only 1,000 or 2,000 videos. Furthermore, the performance difference between 5,000 and 10,000 videos is negligible, suggesting that 5,000 training videos are sufficient for the 3D-aware pretraining stage. Collecting data beyond 10,000 videos appears to offer no significant advantage and is both unnecessary and inefficient.

Additionally, as demonstrated in Fig. A1, the trained model benefits from increased data diversity, where we shortened training steps due to time constraints.

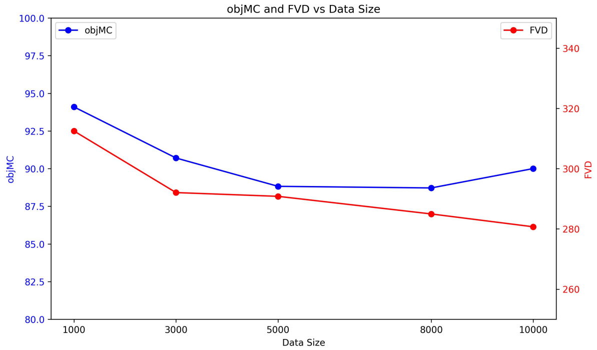

Figure A1. Visualization of objMC and FVD with scaling diversity.

B. More Visualization Results

In this section, we provide more visualization results from our model in Fig. A3, including different rotational trajectories for single-object and multiple-object controlling, and the overall camera controlling.

It can be observed that our model generates both precise rotational and translational motion following trajectories for various objects and also maintains superior object entity and video quality with potential wide-range motions.

C. Visualization Results for Ablation Studies

In this section, we present additional visualization results (Figs. A4 and A5) to complement the metric-based experiments for our ablation studies discussed in the main paper (Sec. 5.2), including the ablation performance on open-

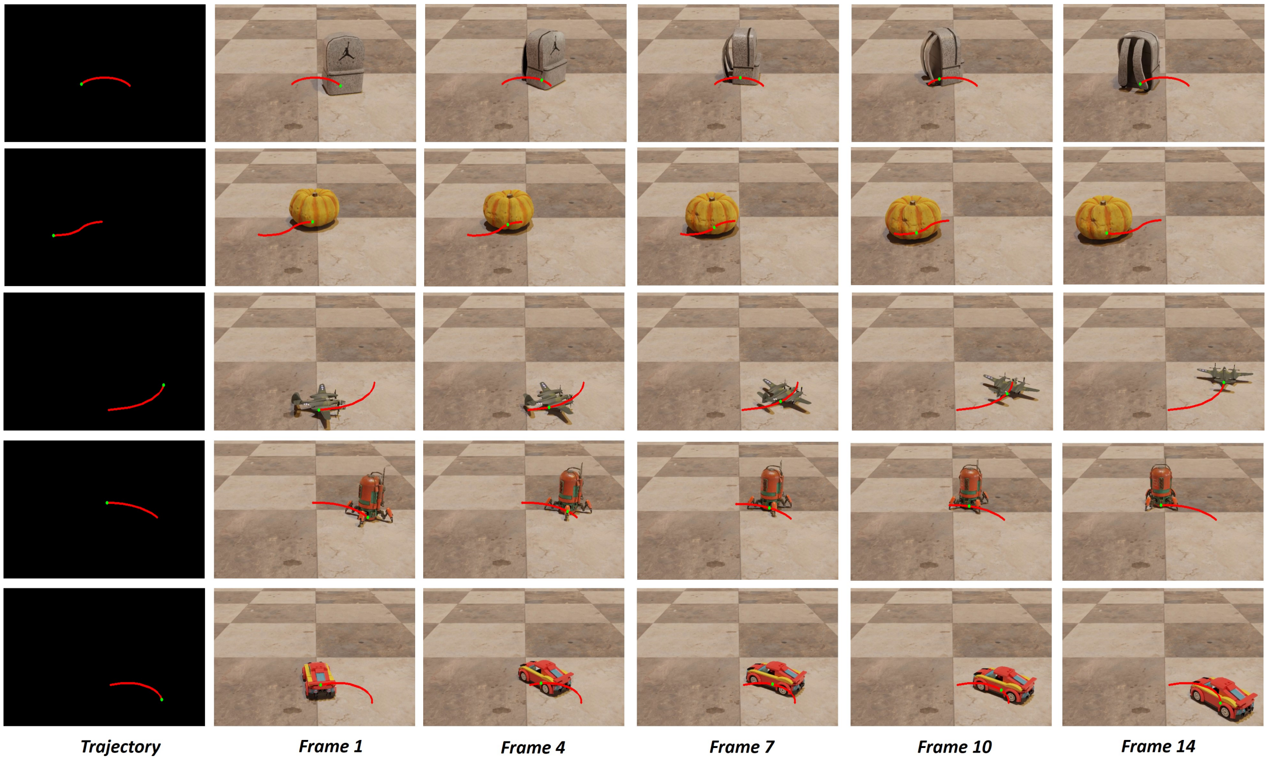

Figure A2. Visualization for several animated samples from our trajectory augmented synthetic dataset.

domain videos and the pretraining stage using synthetic data.

C.1. Performance on Open-Domain Dataset

Fig. A4 shows ablation results on the open-domain dataset. The model trained without two-stage pretraining exhibits poor trajectory-following capability in later frames, demonstrating a lack of temporal consistency. For the model trained without bounding box supervision ('No bbox stage'), the pose changes in the generated rotational motions under wide-range motion scenarios are notably less pronounced compared to the final model. Additionally, removing the spatial enhancement loss during training leads to a collapse in object identity, resulting in poor visual coherence. While the model trained without the camera disentanglement module retains comparable pose-aware generation capability for rotational and accurate motions, it suffers from misaligned camera perspectives and increased instability during inference, leading to frequent camera movements that degrade the overall quality.

C.2. Performance on Synthetic Dataset

We further present additional visualizations to support the ablation study conducted during our designed pretraining stage. As demonstrated in Fig. A5, the model trained without spatial enhancement loss exhibits a notable degradation in trajectory-following accuracy. Additionally, the model that omits first-stage pretraining with 3D bounding boxes experiences significant object collapse in the final few frames, accompanied by corresponding worse motion accuracy.

D. Implementation Details

Training details. Our full training pipeline includes three stages: two-stage pose-aware pre-training on a synthetic dataset and final-stage camera-disentangled finetuning on open-domain videos. All training is deployed on a single A100, requiring about 40G memory usage for the batch size of 1. Each stage of the pre-training took 5k steps using an AdamW optimizer with a 1e-5 learning rate. Our final finetuning took another 10k steps.

Real-world video annotation. To annotate the real-world video with trajectory and camera poses, we exploit two separate steps. For the trajectory, following DragAnything [4], we compute the center locations of objects based on their corresponding instance masks in our selected dataset. We then employ CoTracker2 [2] to extract the motion trajectories of these center points, which serve as conditional input. To ensure consistency between synthetic and real-world data, the extracted trajectories are formatted as discrete pixel space points as in our synthetic dataset. As for the camera pose, we extract camera parameters from videos through DROID-SLAM [3].

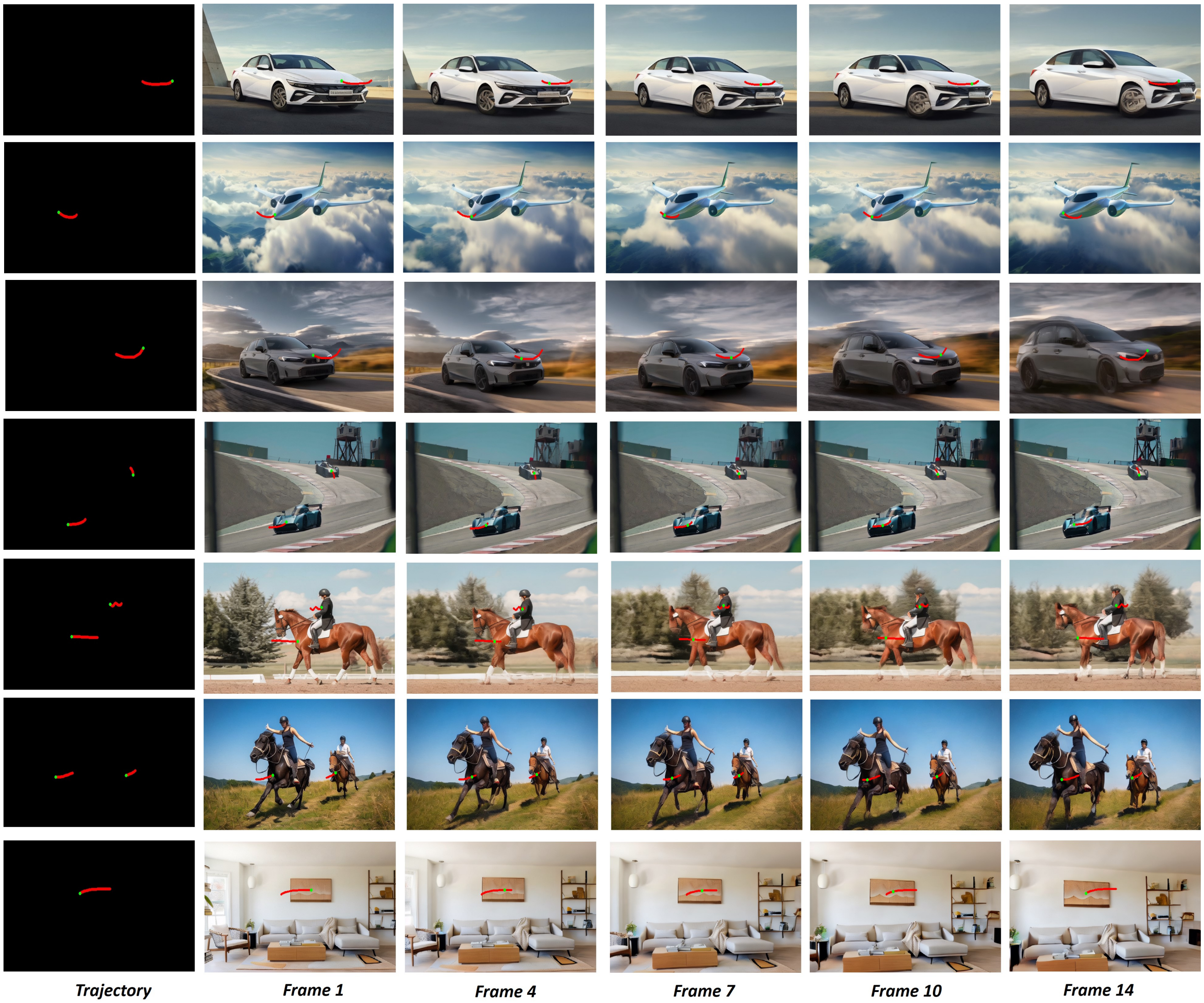

Figure A3. More visualization results of our *PoseTraj* facing various rotational trajectories for single-object and multiple-object, and overall camera controlling.

Trajectory sampler. For enhanced robustness during inference, we modify the original trajectory at the object center by sampling trajectories more sparsely within the projected 2D bounding boxes, with n sampled dragging initial points ($n \leq 8$). These sampled points are dragged along the original trajectory's motion path, and their movements are visualized as images for further training.

E. Discussion

E.1. Why Using 3D Poses as Supervision Signal

In the context of 3D-aligned video generation, various signals, such as depth or depth heatmaps, encode potential 3D information. Compared to using depth as an internal supervision signal, 3D bounding boxes provide explicit object-level pose and approximate location, significantly improving object appearance refinement. Another possible approach is to introduce an additional depth dimension during generation. However, unlike 2D object localization, accurately estimating depth during inference remains highly challenging, often resulting in severe mismatches at the inference stage.

E.2. Limitation and Future Work

Through our experiments, we identified three primary limitations of the current model:

- **Limited capability for wide-range rotations of dynamic objects**. While the model can generate stable and accurate pose-aware rotational motions for *static* objects such as cars, planes, and horses, it struggles with wide-

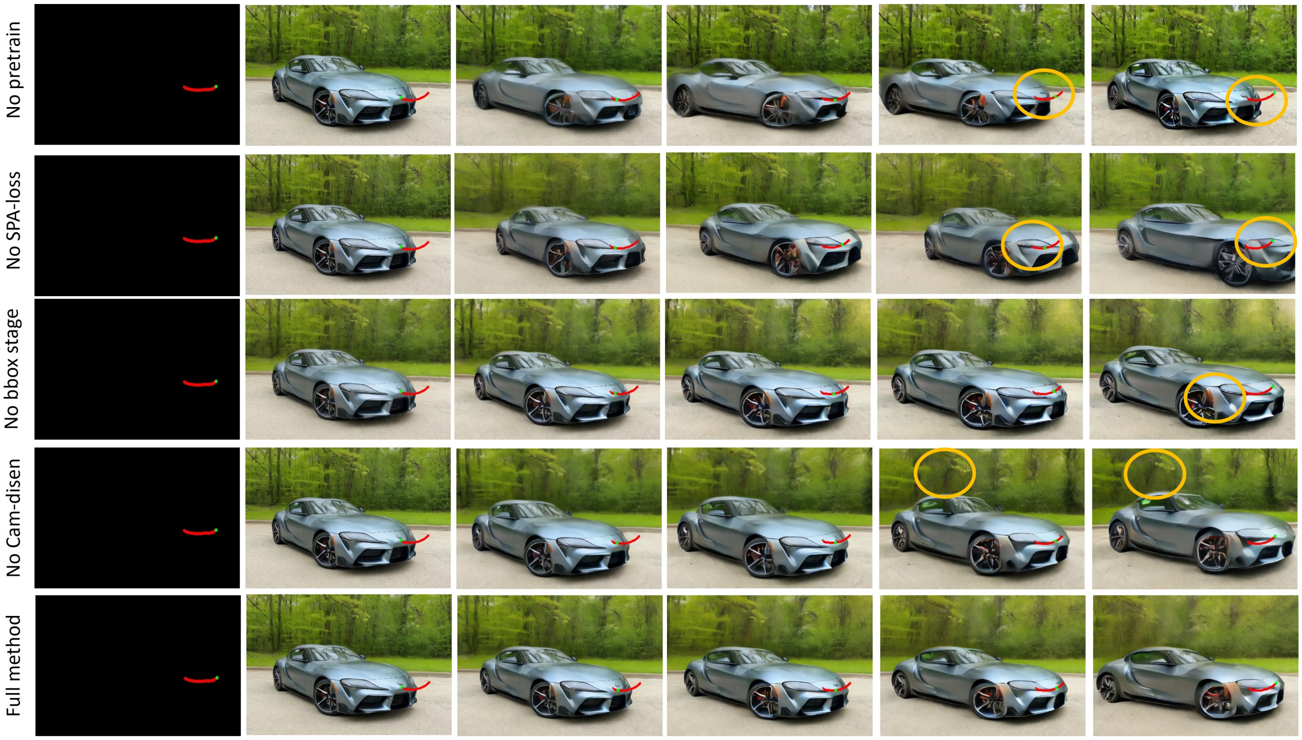

Figure A4. Visualization for generated results of ablation study on open-domain videos.

range rotations for *dynamic* objects, such as humans. This limitation primarily stems from the lack of rotationally dynamic objects, such as people or animals, in the training dataset. A potential solution is to incorporate additional animatable objects, such as walking bears or running avatars, into the synthetic dataset during pretraining.

- **Insufficient camera control capacity**. Despite employing a camera-disentanglement module to enhance object-centric trajectory understanding, the current module fails to provide precise camera control. This issue could be addressed by incorporating large-scale camera-specific datasets, such as RealEstate10K [5], for further training or by directly integrating precisely controlled moving cameras into the synthetic dataset during the pretraining phase.
- **Blurry background in large motions**. Our model may generate a blurry background and this problem comes from the following two aspects. Firstly, the base SVD model struggles to maintain a consistent background when handling large motions. Secondly and most importantly, while our model improves trajectory-matching accuracy for large motions, the inherent blurriness in large movements from training data negatively impacts overall performance. This problem can be mitigated by fine-tuning on high-quality datasets.

References

[1] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[2] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. In *Proc. ECCV*, 2024. 2

[3] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021. 2

[4] Wejia 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. *arXiv preprint arXiv:2403.07420*, 2024. 2

[5] 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

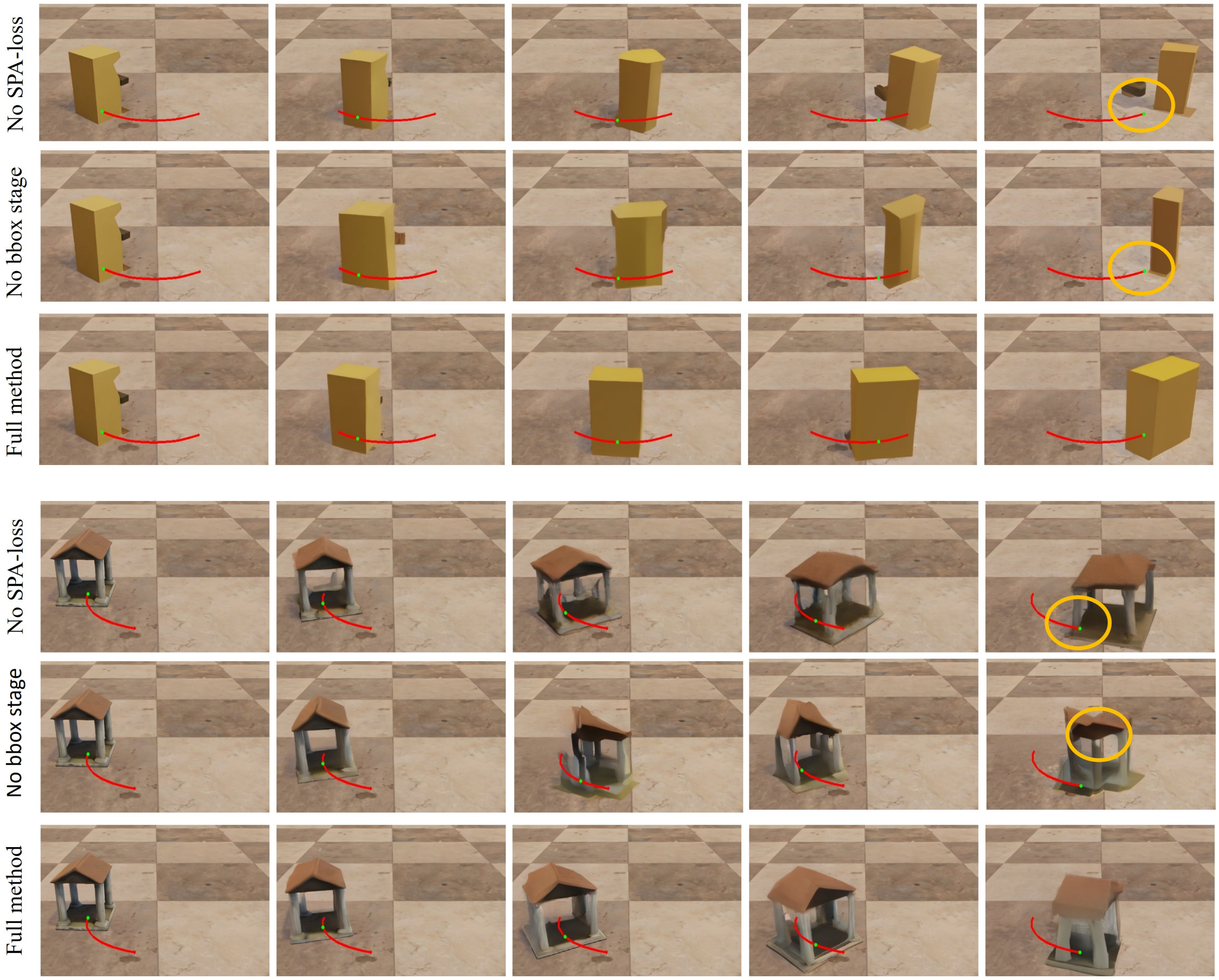

Figure A5. Visualization for the generated results of ablation study on the synthetic dataset.