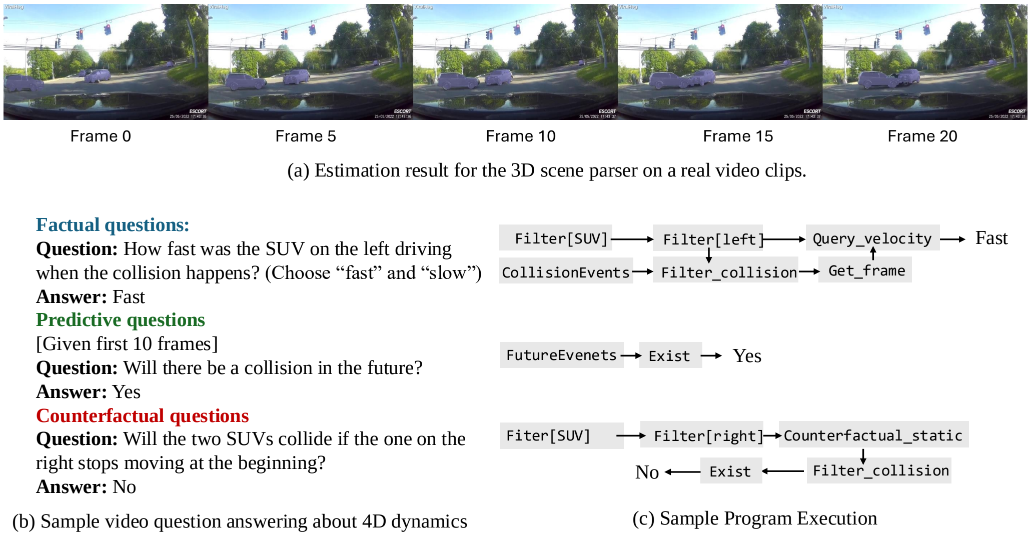
1 Domain extension for the NS-4DPhysics to the real video

We design a case study to demonstrate the real-world generalization ability of the proposed NS-4DPhysics pipeline. Following the architecture of the proposed model, we train the 3D scene parser on the Pascal3D+ dataset, which contains 3D pose annotations but lacks object appearance labels. The qualitative reconstruction results, as shown in Fig. 1(a), demonstrate accurate estimations on the real video, where 4D dynamic properties, including velocities and accelerations, can be effectively inferred. Although the model is not trained on object appearances, its capabilities can be extended by incorporating object classifiers with proper annotations or by enabling open-vocabulary recognition through pretrained large vision-language feature embeddings (e.g., CLIP). This as an important direction in our future work.

Additionally, as shown in Fig. 1(b) and (c), similar types of questions can be posed for the given video, which can then be answered by executing the program step-by-step.



(a) Estimation result for the 3D scene parser on a real video clips.

Factual questions:
Question: How fast was the SUV on the left driving when the collision happens? (Choose “fast” and “slow”)
Answer: Fast

Predictive questions
[Given first 10 frames]
Question: Will there be a collision in the future?
Answer: Yes

Counterfactual questions
Question: Will the two SUVs collide if the one on the right stops moving at the beginning?
Answer: No

(b) Sample video question answering about 4D dynamics

(c) Sample Program Execution

Figure 1: Qualitative reconstruction results from real video data. (a) Estimation results from the 3D scene parser, where 4D dynamic properties, including velocities and accelerations, are effectively inferred. (b) Example questions about the video, which can be answered by executing the corresponding program step-by-step as shown in (c).