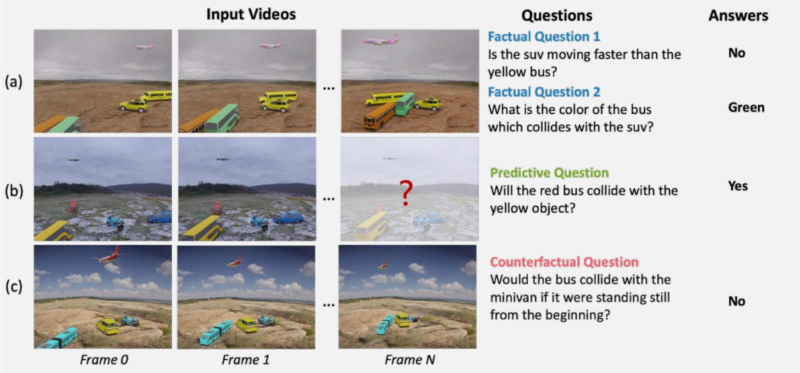
Compositional 4D Dynamic Scenes Understanding with Physics Priors for Video Question Answering

Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, Alan Yuille

New Benchmark: DynSuperCLEVR

A Video question answering dataset over the **4D dynamics** properties of objects (velocity, acceleration) and their interactions (collisions)

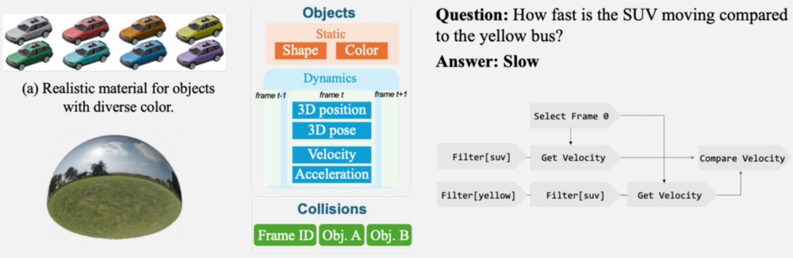


Benchmark Spec.

(a) We use the 21 3D object meshes and generate more realistic textures for different colors.

(b) Annotations includes static and dynamic properties for objects and collision events. Specifically, we study the 3D velocities, accelerations and the collisions in 3D spaces.

(c) Three question types cover the factual questions, predictive questions and counterfactual questions. Each question has new operation programs for 4D dynamic properties reasoning.



(a) Realistic material for objects with diverse color.

(b) Environment map in dome shape as background.

(c) Structure of scene annotations for objects and events.

(d) Example of program execution for questions answering on dynamical properties

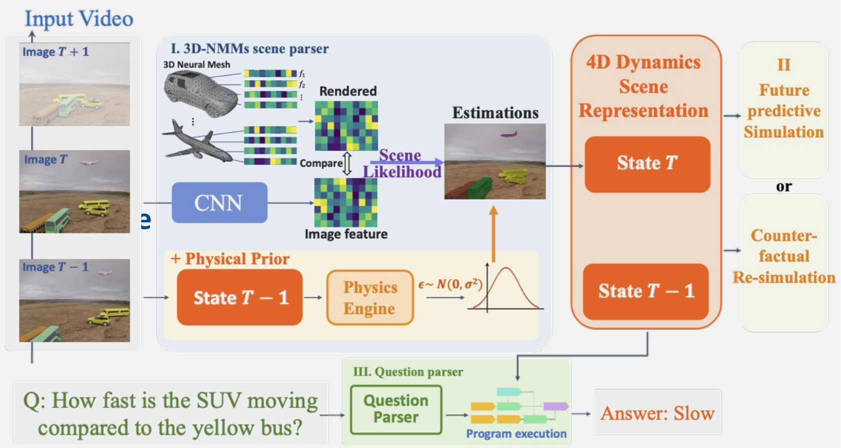
Model: NS-4DPhysics

A neural-symbolic model on 4D scene representation with physical prior. The pipeline is below:

(1) 3D scene parser: Videos ⇨ 4D scene representation

(2) Question parser: Questions ⇨ Programs

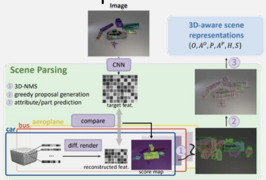
(3) Program executor: 4D scene representations + programs ⇨ Answer



3D scene parser

(a) 3D neural mesh model which use render-and-compare to estimate 6D poses.

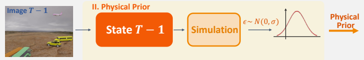
(b) Trained by aligning 3D-rendered features with 2D image features, enabling inference of object pose and category via render-and-compare.



(c) Inference time: Integrating discriminative physical engine output as physical prior

(i) Physical prior compute from previous state

$$(R_t, T_t) \mid (\hat{R}_{t-1}, \hat{T}_{t-1}) \sim \mathcal{N}(\mathrm{PE}(\hat{R}_{t-1}, \hat{T}_{t-1}), \sigma^2 I)$$



(ii) Neural mesh model with differentiable rendering

$$\hat{R}_t, \hat{T}_t = \arg\max_{R_t, T_t} p(F_t \mid O_c, R_t, T_t, B)$$

(iii) Analysis by synthesis

$$\hat{R}_t, \hat{T}_t = \arg\max_{R_t, T_t} p(F_t \mid O_c, R_t, T_t, B) \cdot q(R_t, T_t \mid \hat{R}_{t-1}, \hat{T}_{t-1}).$$

Benchmark Results

Performance on the *DynSuperCLEVR* testing split for each question type: factual, predictive, and counterfactual.
Factual questions are further divided into sub-types: **Vel**ocity, **Acc**eleration, and **Col**lision, with "**All**" representing overall accuracy. The average is taken as the overall accuracy across the three question types.†

	Average	Factual All	Factual Vel.	Factual Acc.	Factual Col.	Predictive	Counterfactual
CNN+LSTM	48.03	40.63	41.71	56.79	25.37	56.04	47.42
FiLM (Perez et al., 2018)	50.18	44.07	48.58	53.09	26.87	54.94	51.54
NS-DR (Yi et al., 2019)	51.44	51.44	55.63	46.34	46.86	-	-
PO3D-VQA (Wang et al., 2024)	<u>62.93</u>	<u>61.22</u>	<u>62.21</u>	<u>73.17</u>	<u>51.20</u>	65.33	62.24
InternVideo (Wang et al., 2022)	52.62	51.07	59.29	49.08	36.06	54.74	59.18
Video-LLaVA † (Lin et al., 2023)	38.09	37.04	37.62	52.76	23.56	38.78	40.88
PLLaVA † (Xu et al., 2024)	59.24	54.61	55.00	63.80	46.63	<u>67.52</u>	<u>73.47</u>
GPT-4o†	51.59	50.82	51.19	57.67	44.71	54.38	50.00
GPT-4o + reasoning †	56.06	55.50	58.81	57.67	47.12	56.93	58.16
NS-4DPhysics	**82.64**	**87.70**	**88.66**	**83.73**	**88.46**	**85.71**	**74.51**

Ablation Study

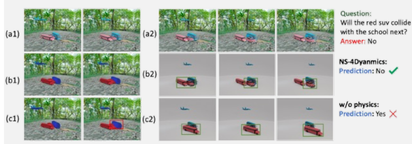
(1) Compare 4D Representation: **with** or **without** physical prior) :(i) and (ii)

(2) Compare Reasoning strategy: symbolic reasoning and GPT4o (i) and (iii)

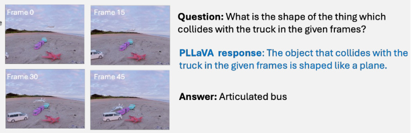
	Average	Factual All	Factual Vel.	Factual Acc.	Factual Col.	Predictive	Counterfactual
4D Representation + SR (Ours)	**82.64**	**87.70**	**88.66**	**83.73**	**88.46**	**85.71**	**74.51**
w/o Physics Prior + SR	75.97	79.68	81.40	81.30	74.88	78.83	69.39
4D Representation + GPT-4o	61.39	65.49	66.67	54.60	71.63	57.14	51.09
Video + GPT-4o	56.06	55.50	58.81	57.67	47.12	56.93	58.16

Visualization

Predictive question



PLLaVA failure example



(b) Misunderstand the collisions