Peekaboo: Interactive Video Generation via Masked-Diffusion

Yash Jain[1]*, Anshul Nasery[2]*, Vibhav Vineet[1], Harkirat Behl[1]

CVPR JUNE 17-21, 2024 SEATTLE, WA

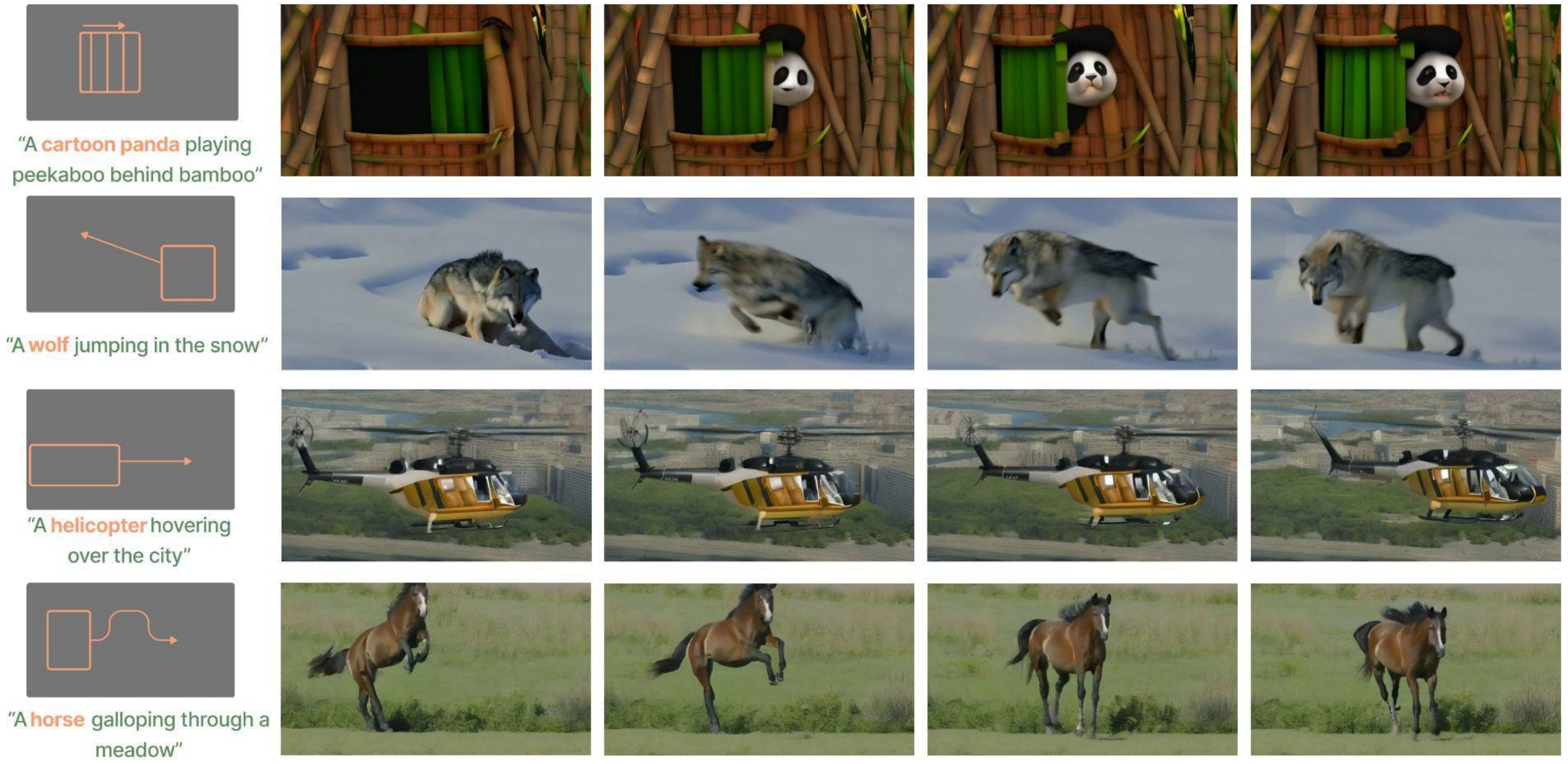
Motivation:

- Text-to-video models are great, but can they control-
  - ○ Size of objects?
  - ○ Path of Motion?
  - ○ Position of objects?
- Only text input is inadequate
- Bounding Box inputs are a good modality for rich description.
- How do we equip existing models to use this input information?

Contributions:

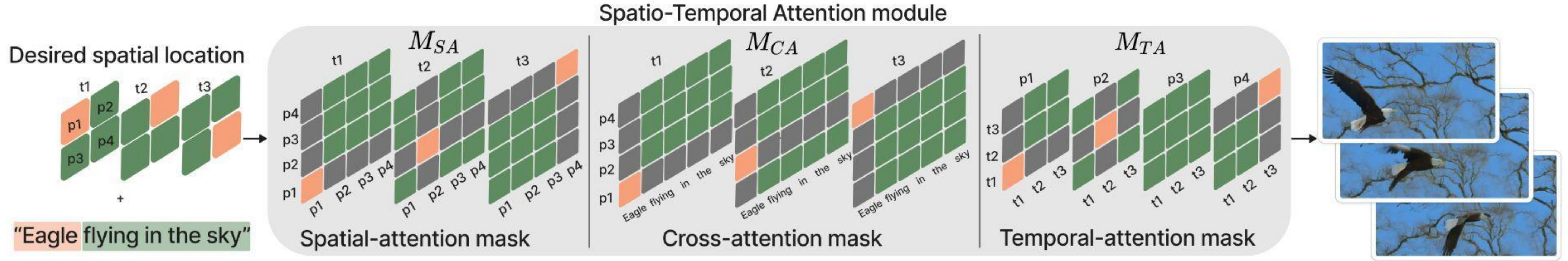
- Peekaboo allows you to **gain interactive control** on videos **-**
  - ○ **On *any* off-the-shelf model.**
  - ○ **Without *any* training.**
  - ○ **With *no* inference overheads.**
- We introduce a new benchmark for controllable video generation
  - ○ Over **800** caption-bbox pairs
  - ○ With 4 different metrics for evaluating performance

Qualitative Results:

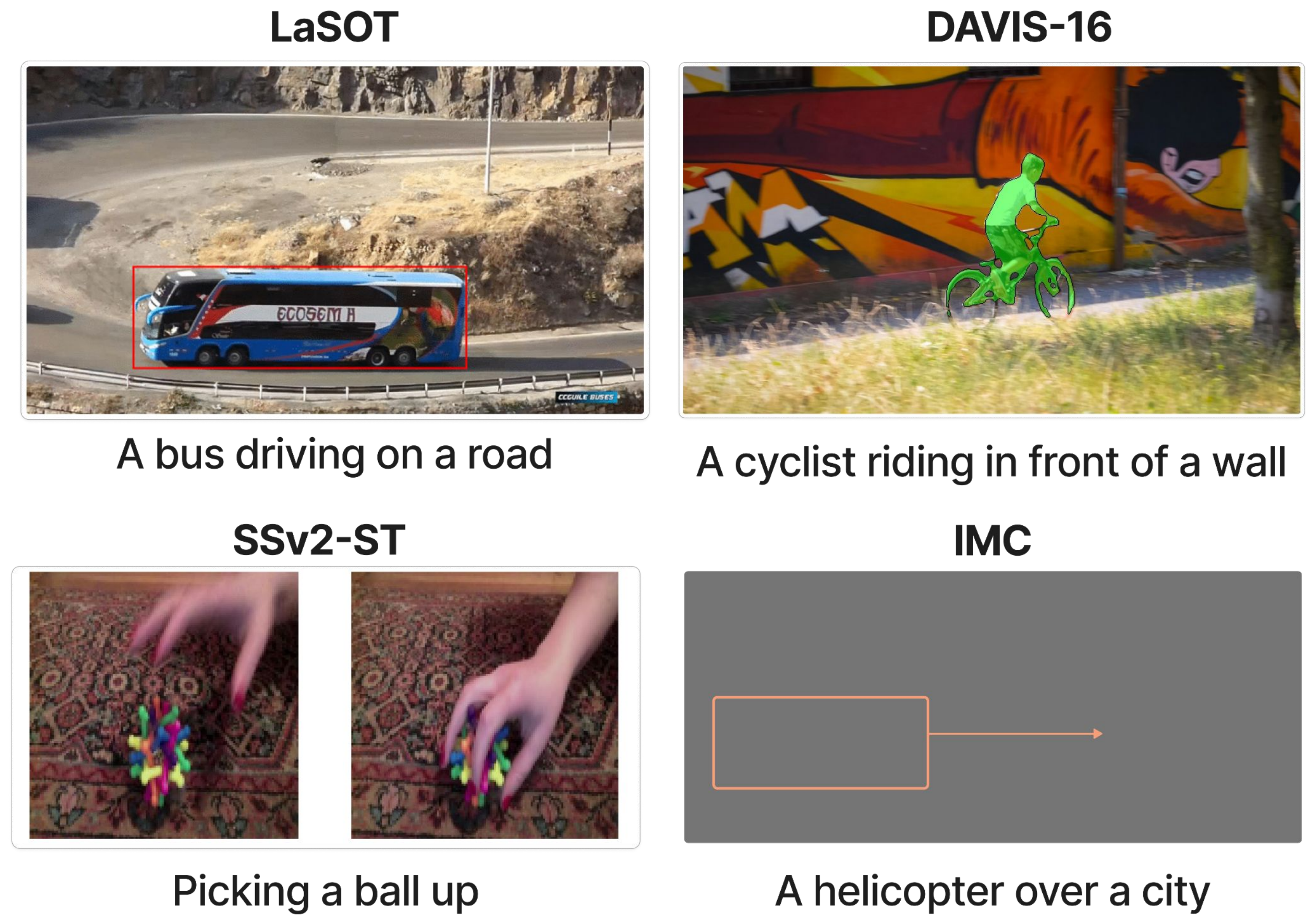

"A cartoon panda playing peekaboo behind bamboo"
"A wolf jumping in the snow"
"A helicopter hovering over the city"
"A horse galloping through a meadow"

Method - Attention (masks) are all you need:

Peekaboo controls outputs by modifying attention masks to the model for some steps.

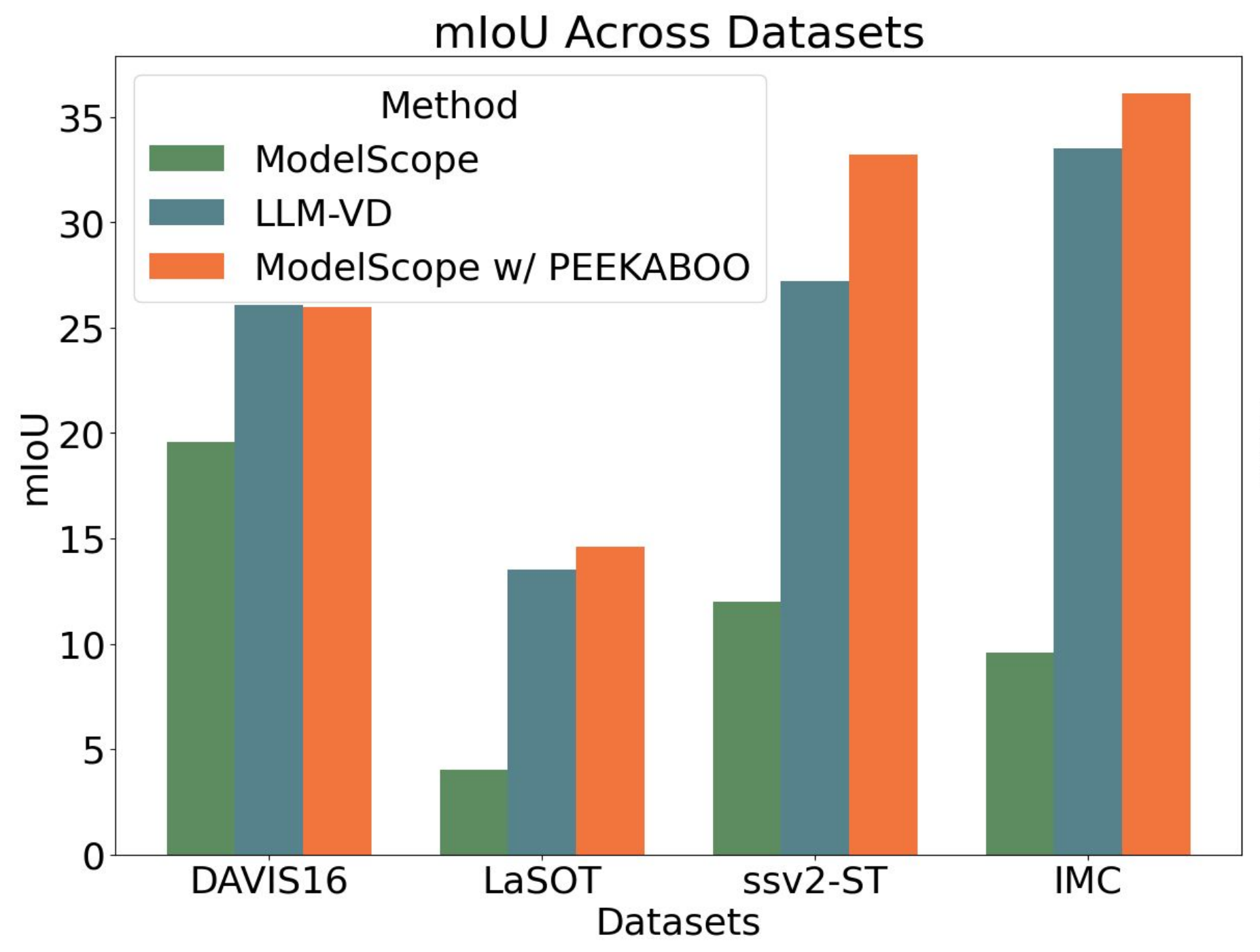


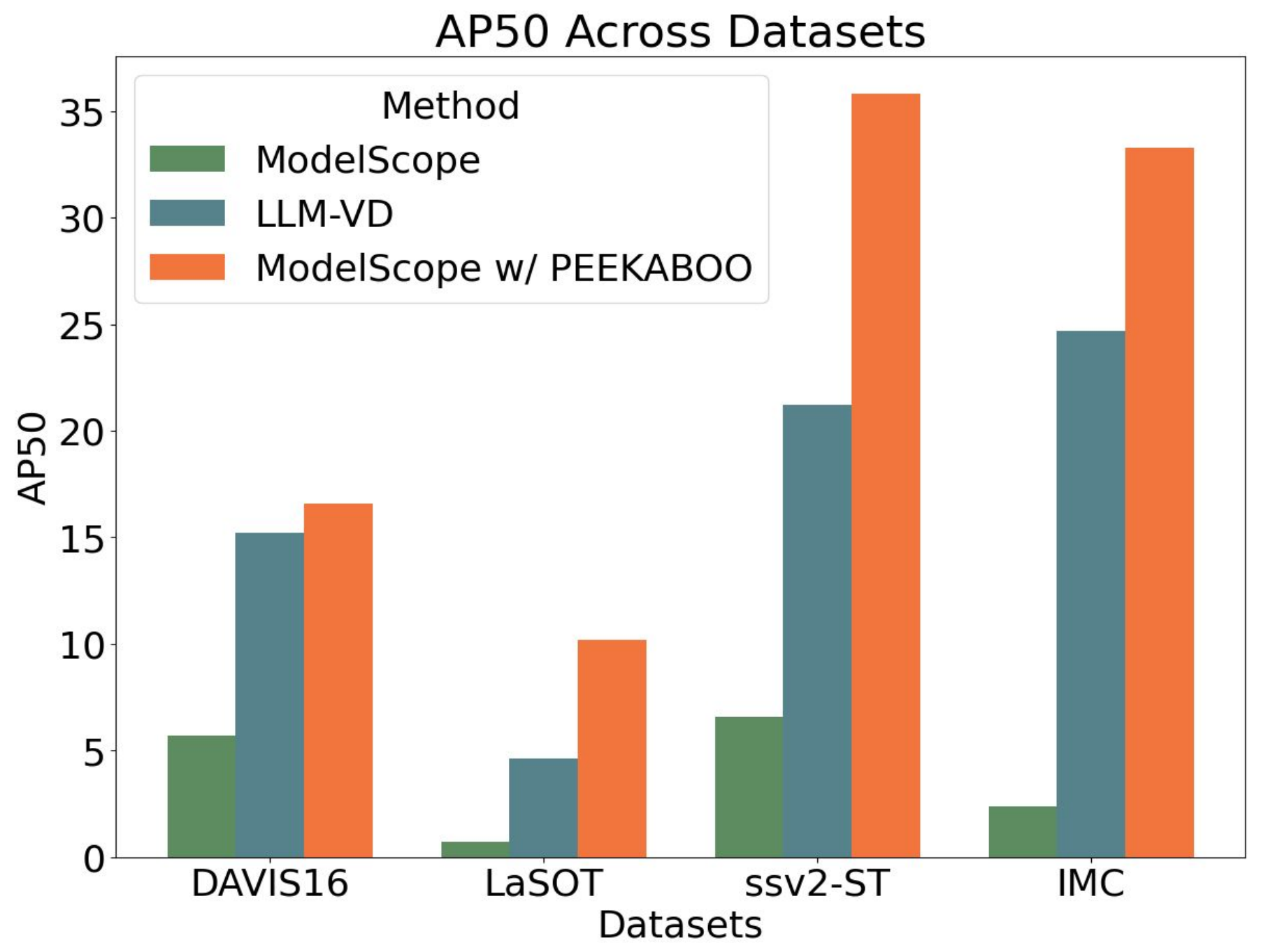
A New Benchmark:



- We repurpose existing video datasets, and also create a new dataset for evaluating methods.
- Metrics: The faithfulness of the generations is measured using AP50, mIoU scores.

Quantitative Results:


mIoU Across Datasets
Method
ModelScope
LLM-VD
ModelScope w/ PEEKABOO
mIoU
DAVIS16
LaSOT
ssv2-ST
IMC
Datasets


AP50 Across Datasets
Method
ModelScope
LLM-VD
ModelScope w/ PEEKABOO
AP50
DAVIS16
LaSOT
ssv2-ST
IMC
Datasets