LLM-grounded Video Diffusion Models

Long Lian[1*], Baifeng Shi[1*], Adam Yala[1,2†], Trevor Darrell[1†], Boyi Li[1†]

[1]UC Berkeley [2]UCSF *Equal contribution †Equal advising

ICLR 2024

Our code is available at llm-grounded-video-diffusion.github.io

Text-to-video diffusion models can generate realistic videos

Yet they still suffer from insufficient prompt understanding

Prompt: A raccoon on a wooden barrel floating on a river

The raccoon is not on a wooden barrel?

— ZeroScope (baseline)

Prompt: A brown bear dancing with a yellow pikachu

The bear blends in with the pikachu?

— ZeroScope (baseline)

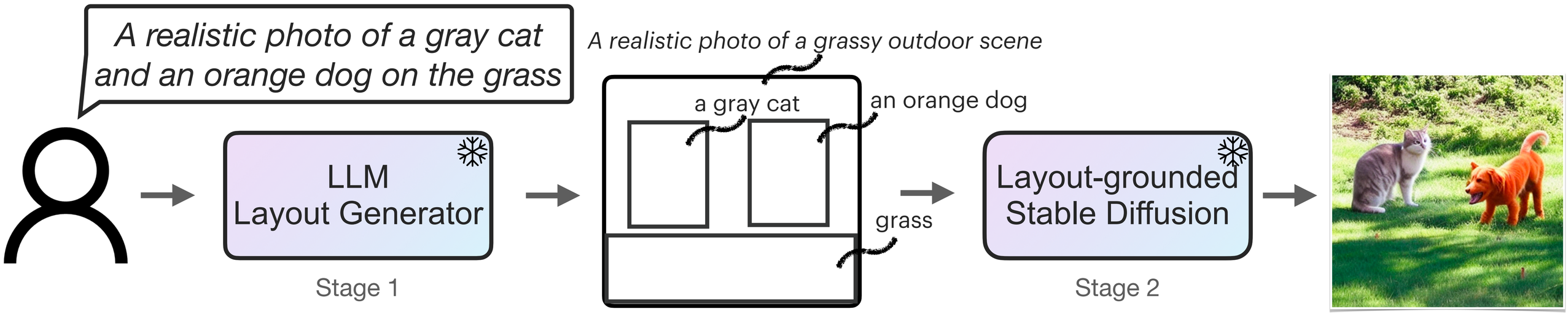
Prompt: A bird flying from the left to the right

The bird is flying towards the left?

— ZeroScope (baseline)

Improve the prompt understanding capabilities of text-to-video diffusion models **without** fine-tuning by introducing LLMs for grounding

Preliminaries: LLM-grounded Diffusion (LMD)

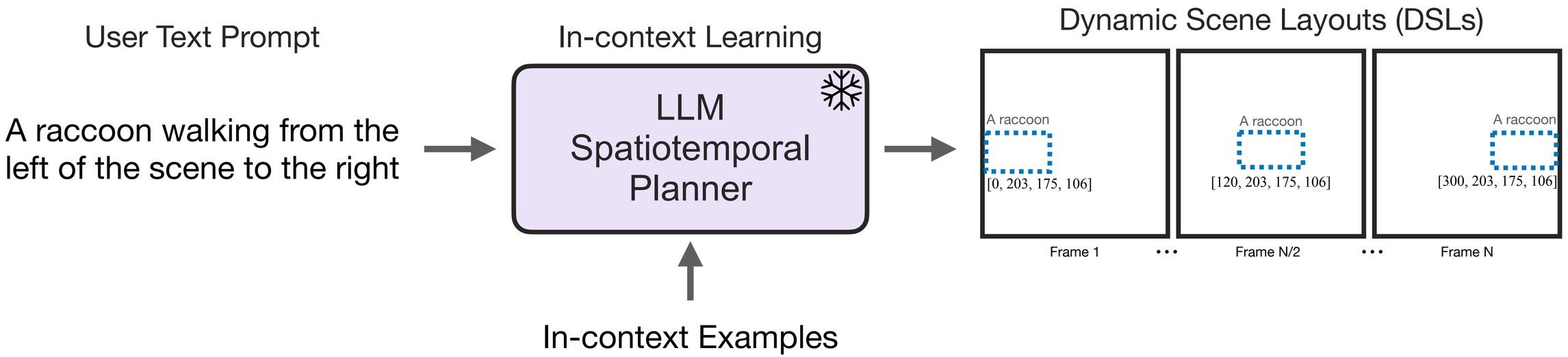


L. Lian, et al. *LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models*. TMLR 2024.

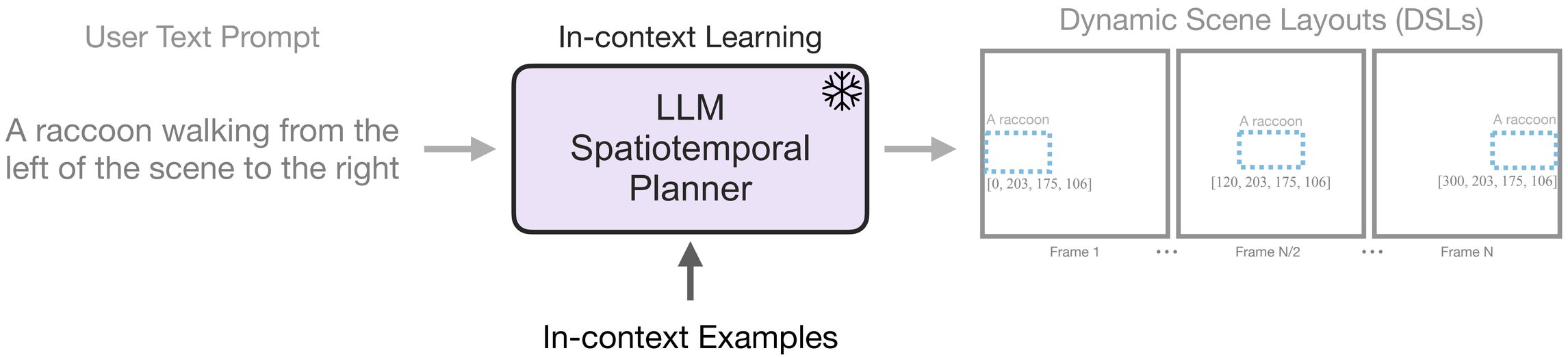
We know that LLMs can reason about spatial layouts

Do they understand **motion** as well?

Do LLMs Understand Spatiotemporal Dynamics?



Designing In-context Examples

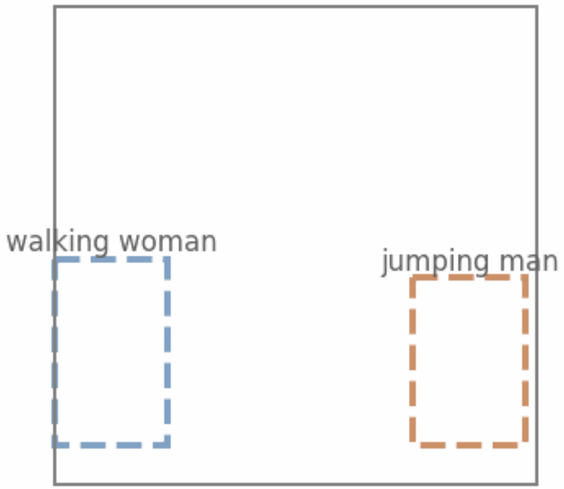


Gravity?

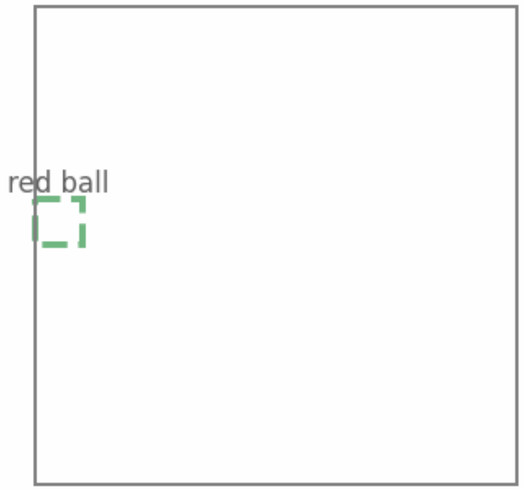
Elasticity?

Perspective projection?

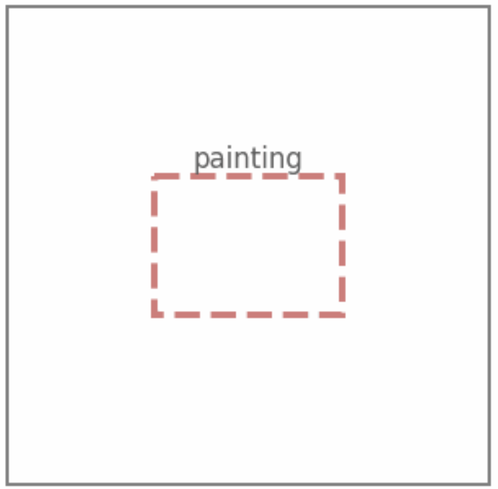
Demonstrating Desirable Properties to LLMs



In-context example for **gravity**: *A woman walking from the left to the right and a man jumping on the right in a room*

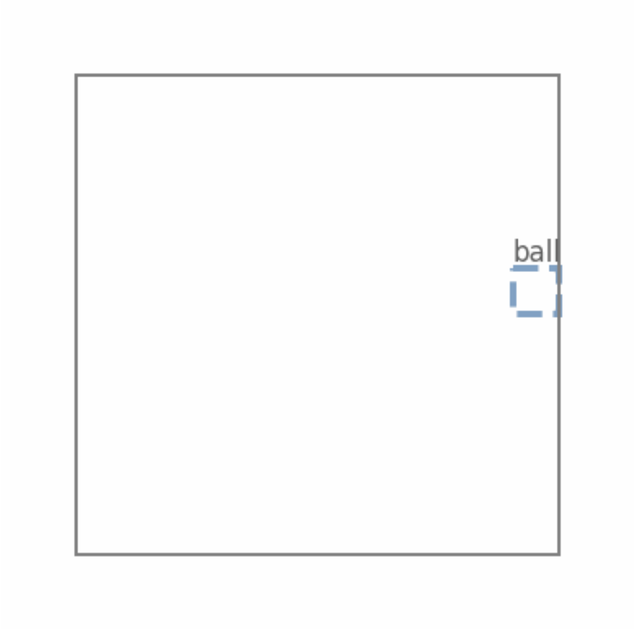


In-context example for **elasticity**: *A red ball is thrown from the left to the right in a garden*



In-context example for **perspective camera projection**: *The camera is moving away from a painting*

Surprising Generalization Capabilities to Unseen Properties



LLM output: *A **ball** is thrown out from the right*

Balls bounce ✓

Surprising Generalization Capabilities to Unseen Properties

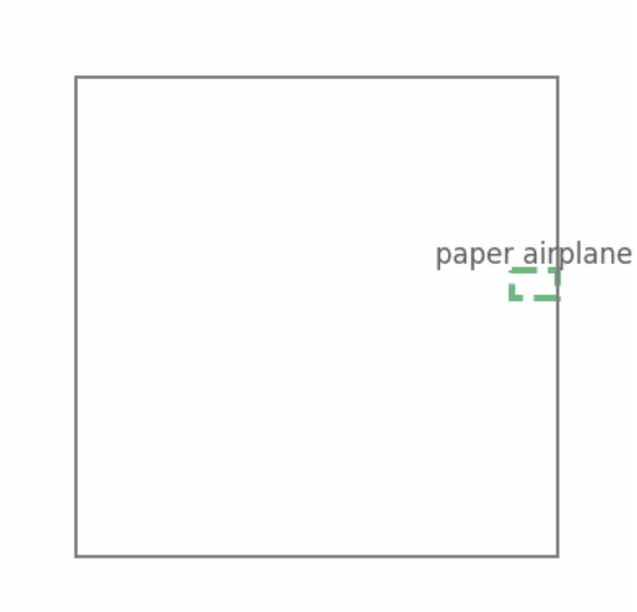
LLM output: *A **rock** is thrown out from the right*

Rocks don’t bounce ✓

(not mentioned in the prompt)

Ability comes from the weights, not instructions or examples!

Surprising Generalization Capabilities to Unseen Properties



LLM output: *A **paper airplane** is thrown out from the right*

Paper airplanes glides ✓

(not mentioned in the prompt)

Air friction is considered!

Surprising Generalization Capabilities to Unseen Properties

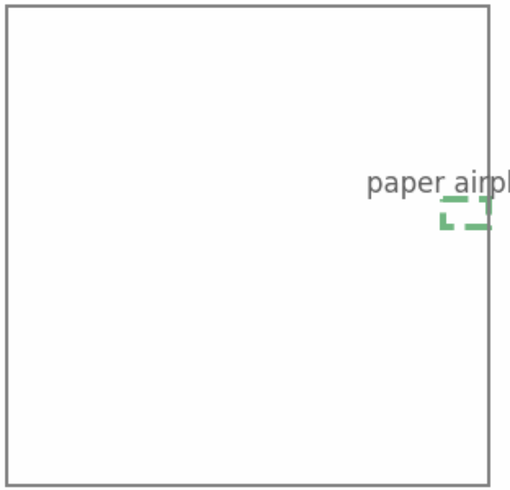
LLM output: *A **ball** is thrown out from the right*

Balls bounce ✓

LLM output: *A **rock** is thrown out from the right*

Rocks don’t bounce ✓

(not mentioned in the prompt)



LLM output: *A **paper airplane** is thrown out from the right*

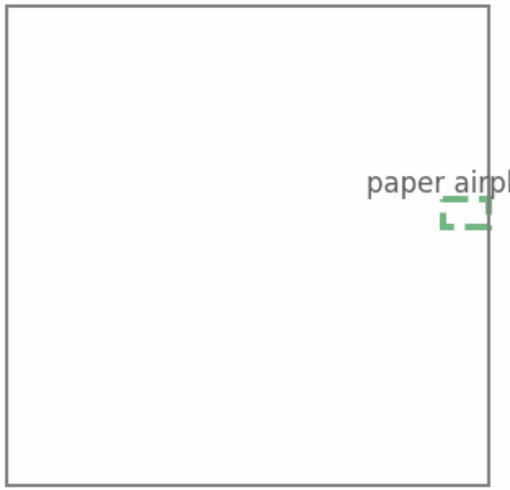
Paper airplanes glides ✓

(not mentioned in the prompt)

Surprising Generalization Capabilities to Unseen Properties

LLM output: *A **ball** is thrown out from the right*

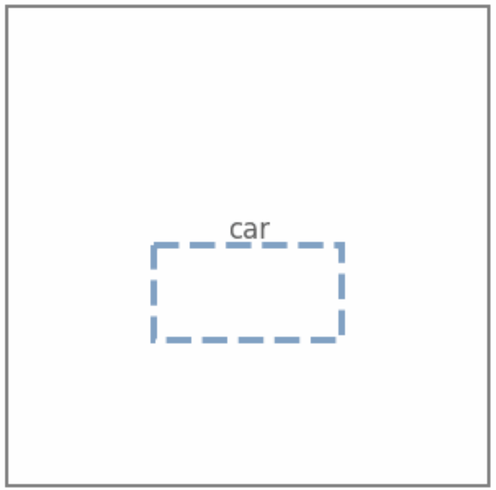
LLM output: *A **rock** is thrown out from the right*



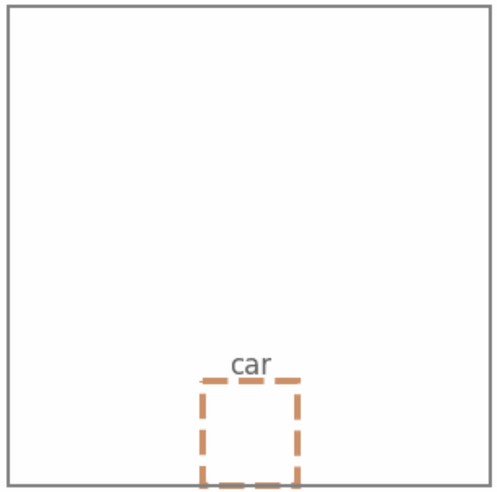
LLM output: *A **paper airplane** is thrown out from the right*

No exhaustive examples needed! 😁

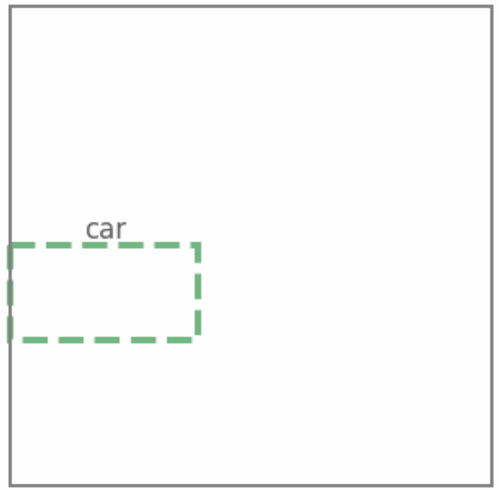
Generalizing to Unseen Viewpoints



LLM output: *A car viewed from the **back** is driving forward*



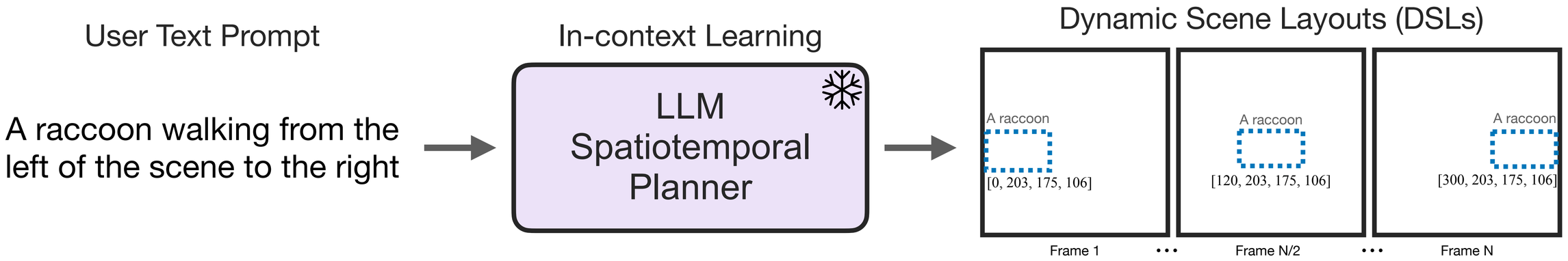
LLM output: *A car viewed from the **top** is driving forward*



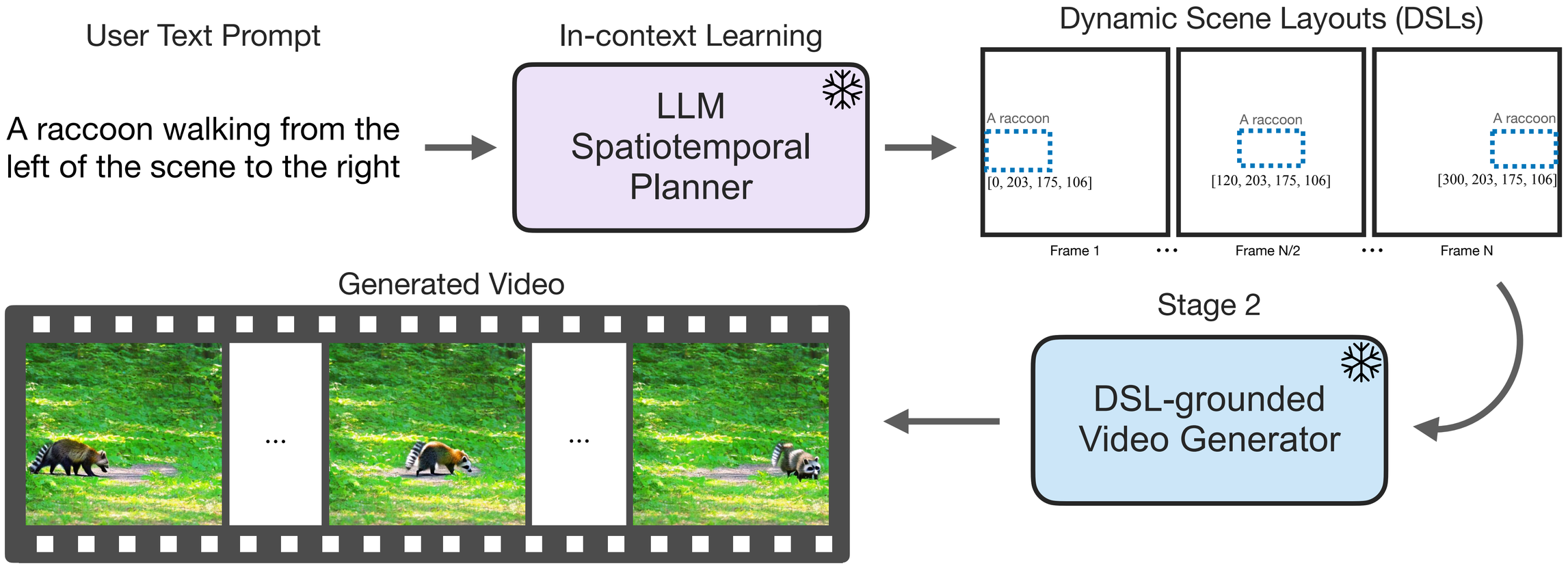
LLM output: *A car viewed from the **side** is driving forward*

Using LLMs to enhance the prompt understanding capabilities of text-to-video diffusion models

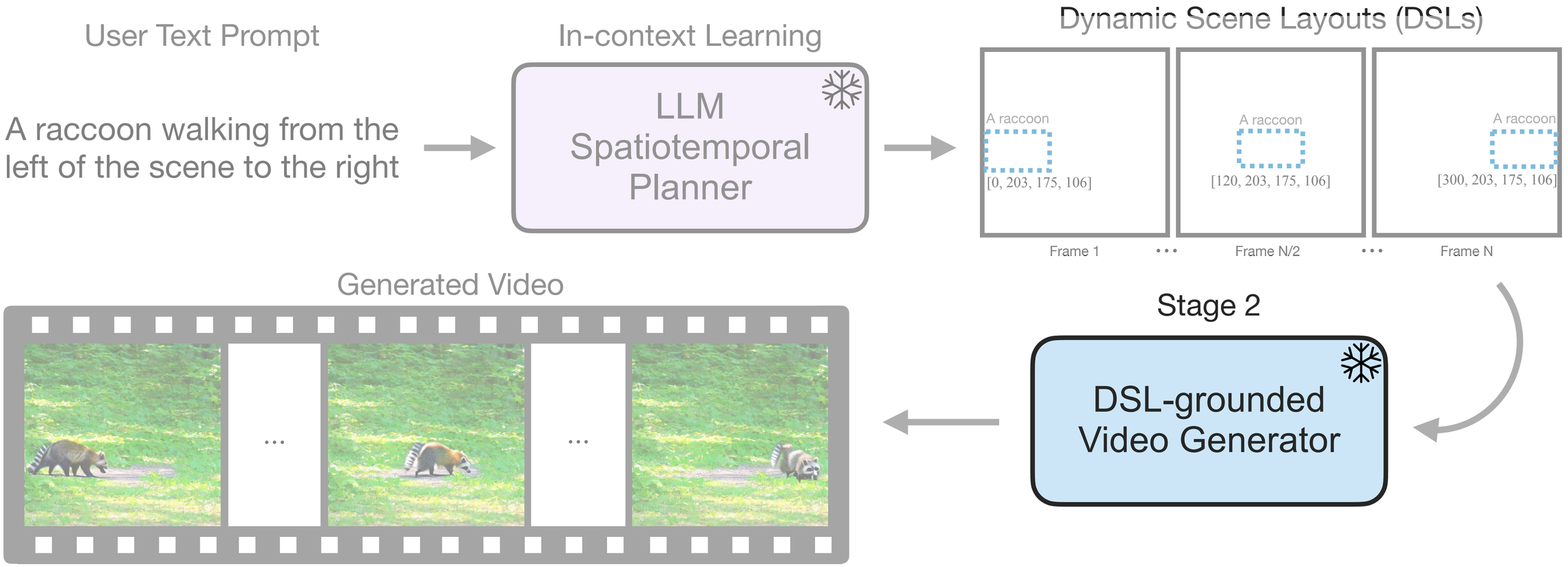
Using LLMs to Enhance Text-to-Video Diffusion Models



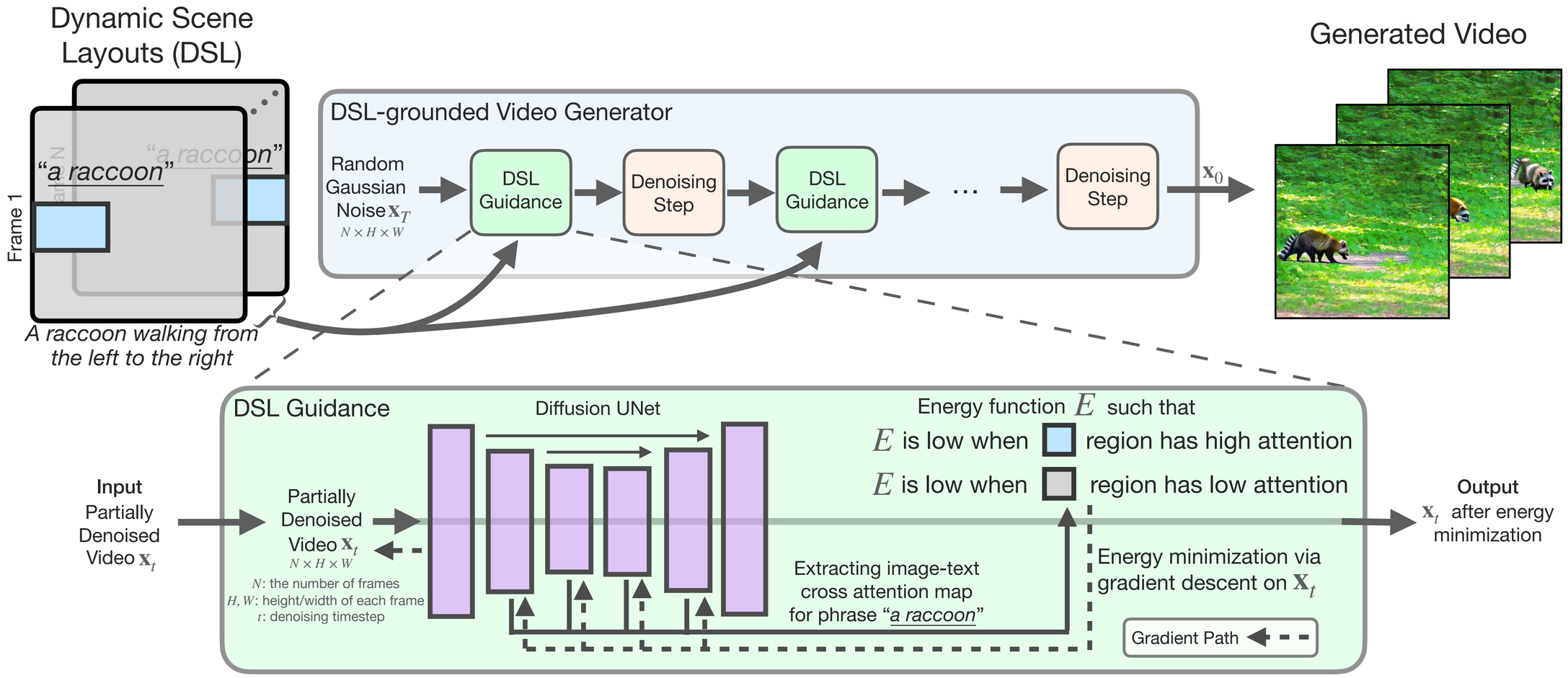
Using LLMs to Enhance Text-to-Video Diffusion Models



Using LLMs to Enhance Text-to-Video Diffusion Models



DSL-grounded Video Generator



A raccoon on a wooden barrel floating on a river

ModelScope (Baseline)

Raccoon not on the barrel ❌

LVD **(Ours)**

Spatial relationships ✓

A brown bear dancing with a yellow pikachu

ModelScope (Baseline)

Mixing pikachu and bear ❌

LVD **(Ours)**

Attribute Binding ✓

A bird flying from the left to the right (of the scene)

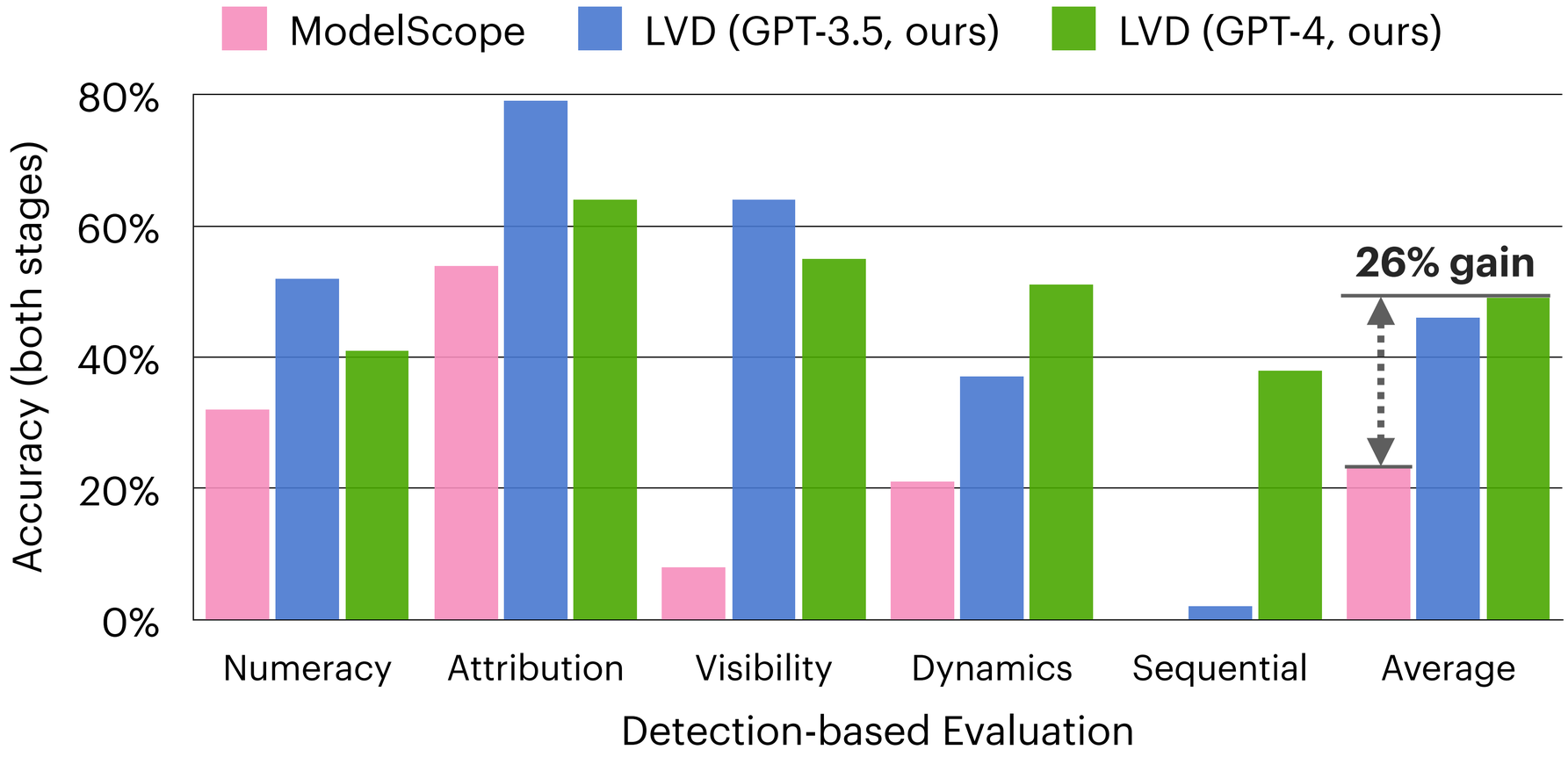
ModelScope (Baseline)

Incorrect flying direction ❌

LVD **(Ours)**

Temporal dynamics ✓

LVD Improves Text-Video Alignment



Core idea:
Text ⇨ Video

Core idea:

Text ⇨ Dynamic Scene Layouts ⇨ Video

LLM-grounded Video Diffusion Models

Long Lian[1*], Baifeng Shi[1*], Adam Yala[1,2†], Trevor Darrell[1†], Boyi Li[1†]

[1]UC Berkeley [2]UCSF *Equal contribution †Equal advising

ICLR 2024

Our code is available at llm-grounded-video-diffusion.github.io