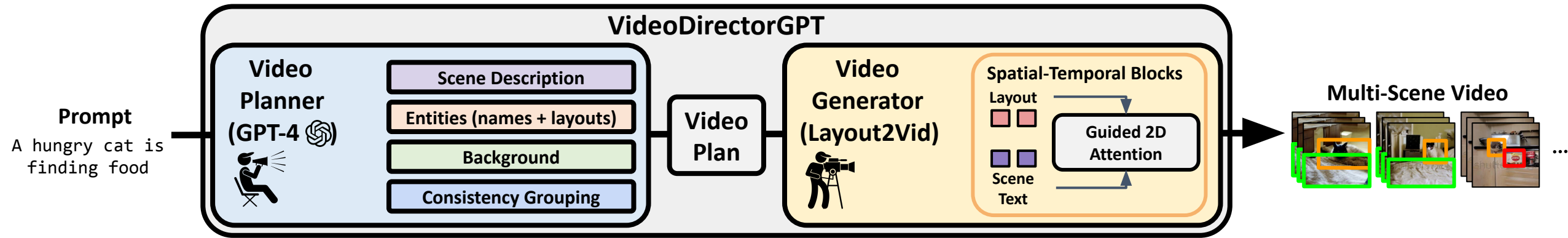


# VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning



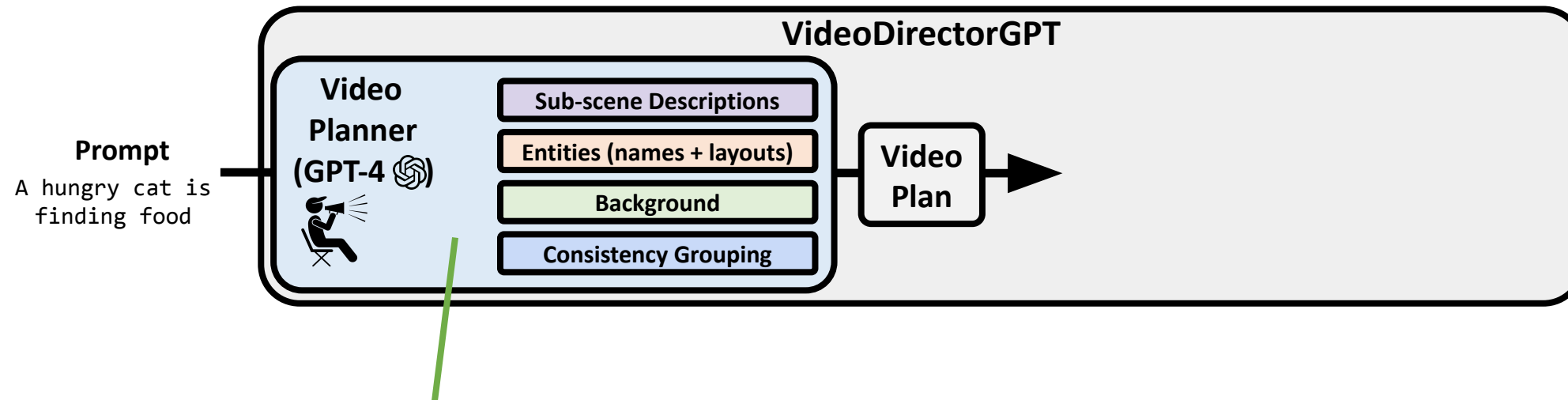
# VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning

## Prompt

A hungry cat is  
finding food

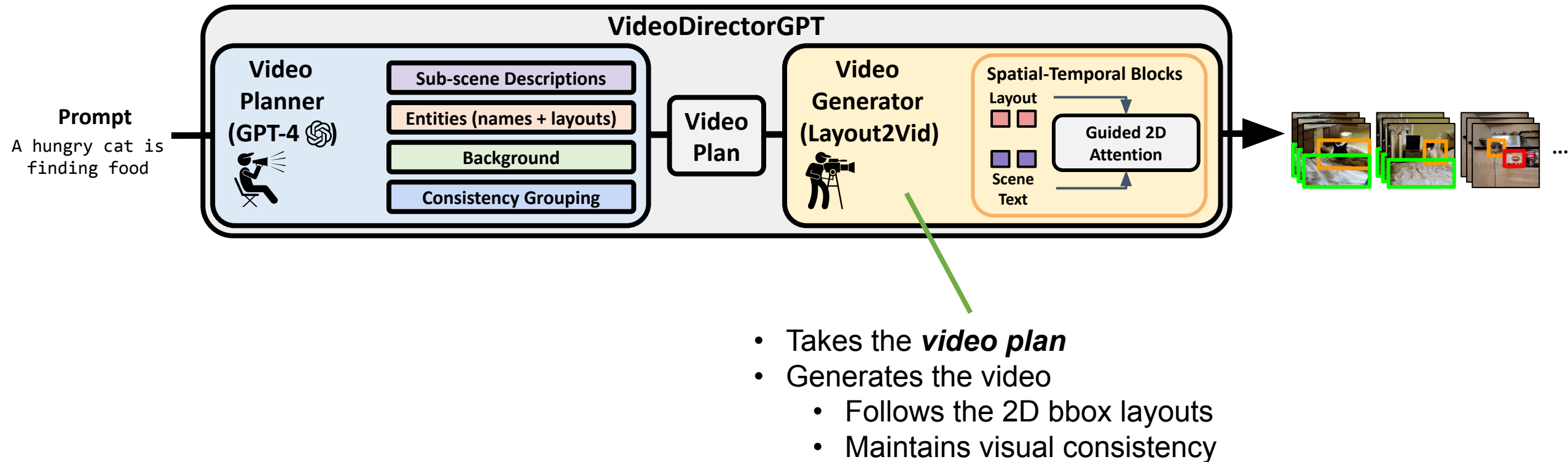
Single input text prompt

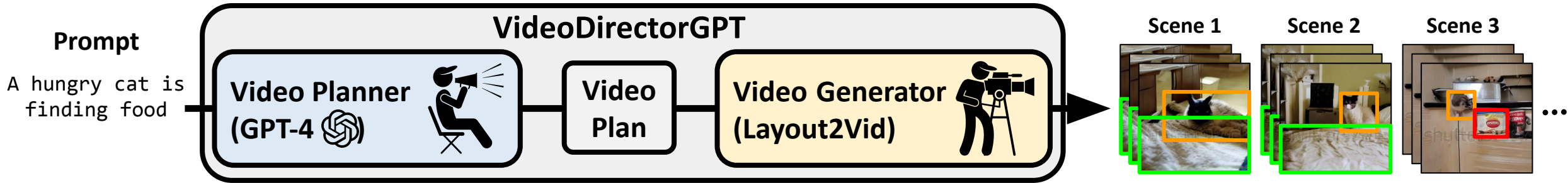
# VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning




- An LLM (GPT-4) creates a **video plan**
  - Sub-scene descriptions
  - Entities (names + 2D bbox layouts)
  - Backgrounds
  - Consistency groupings.

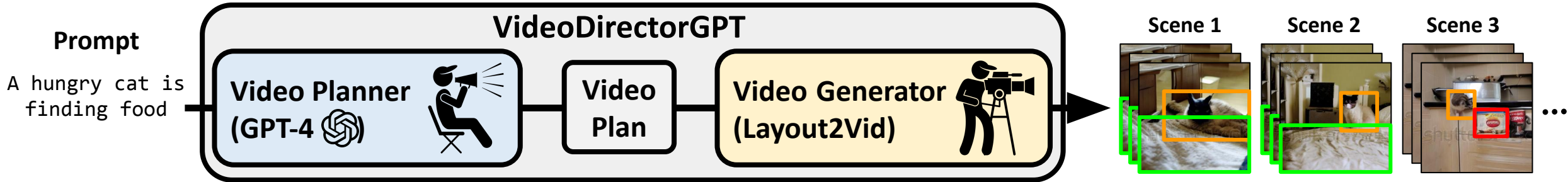
# VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning





Video Planner

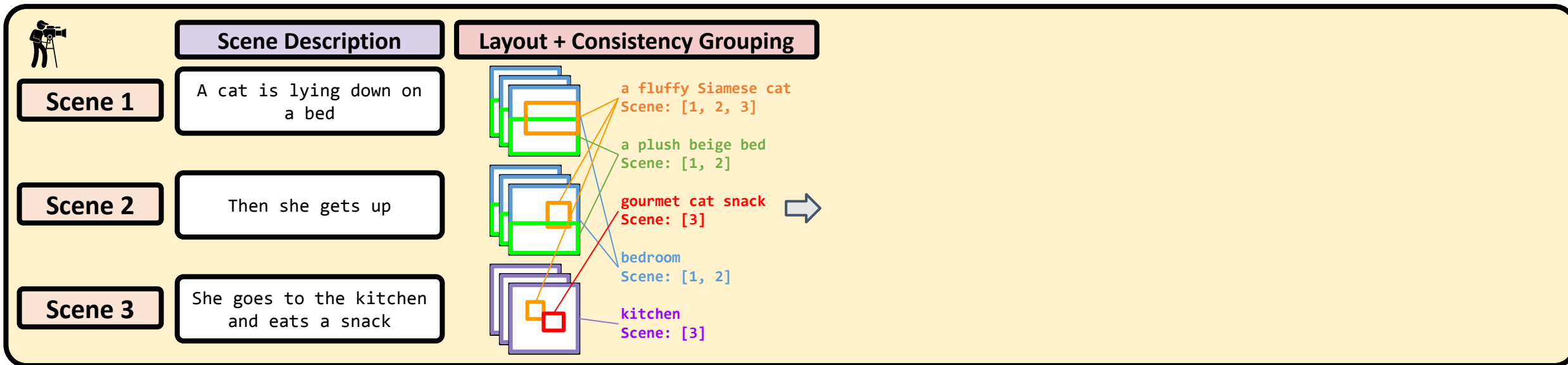
	Scene Description	Entities (names + layouts) with Consistency Grouping	Background
Scene 1	A cat is lying down on a bed	Frame 1: {'a fluffy Siamese cat': [0.25, 0.25, 1.00, 0.70], 'a plush beige bed': [0.00, 0.50, 1.00, 1.00]} Frame 2: {'a fluffy Siamese cat': [0.25, 0.25, 1.00, 0.70], 'a plush beige bed': [0.00, 0.50, 1.00, 1.00]} ...	Bedroom
Scene 2	Then she gets up	Frame 1: {'a fluffy Siamese cat': [0.55, 0.25, 0.85, 0.55], 'a plush beige bed': [0.00, 0.60, 1.00, 1.00]} Frame 2: {'a fluffy Siamese cat': [0.50, 0.30, 0.80, 0.60], 'a plush beige bed': [0.00, 0.60, 1.00, 1.00]} ...	Bedroom
Scene 3	She goes to the kitchen and eats a snack	Frame 1: {'a fluffy Siamese cat': [0.15, 0.20, 0.40, 0.45], 'gourmet cat snack': [0.50, 0.45, 0.80, 0.65]} Frame 2: {'a fluffy Siamese cat': [0.35, 0.30, 0.60, 0.55], 'gourmet cat snack': [0.50, 0.45, 0.80, 0.65]} ...	Kitchen

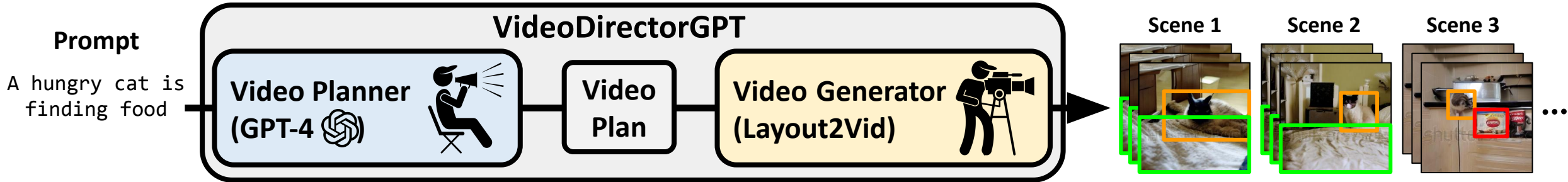


Video Planner

	Scene Description	Entities (names + layouts) with Consistency Grouping	Background
<b>Scene 1</b>	A cat is lying down on a bed	Frame 1: {'a fluffy Siamese cat': [0.25, 0.25, 1.00, 0.70], 'a plush beige bed': [0.00, 0.50, 1.00, 1.00]} Frame 2: {'a fluffy Siamese cat': [0.25, 0.25, 1.00, 0.70], 'a plush beige bed': [0.00, 0.50, 1.00, 1.00]} ...	Bedroom
<b>Scene 2</b>	Then she gets up	Frame 1: {'a fluffy Siamese cat': [0.55, 0.25, 0.85, 0.55], 'a plush beige bed': [0.00, 0.60, 1.00, 1.00]} Frame 2: {'a fluffy Siamese cat': [0.50, 0.30, 0.80, 0.60], 'a plush beige bed': [0.00, 0.60, 1.00, 1.00]} ...	Bedroom
<b>Scene 3</b>	She goes to the kitchen and eats a snack	Frame 1: {'a fluffy Siamese cat': [0.15, 0.20, 0.40, 0.45], 'gourmet cat snack': [0.50, 0.45, 0.80, 0.65]} Frame 2: {'a fluffy Siamese cat': [0.35, 0.30, 0.60, 0.55], 'gourmet cat snack': [0.50, 0.45, 0.80, 0.65]} ...	Kitchen

Video Generator

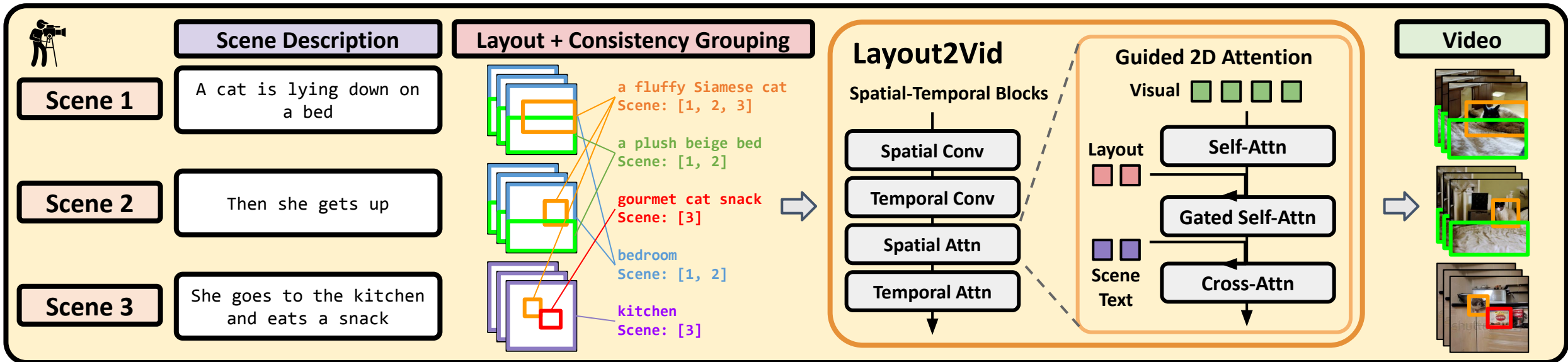




Video Planner

	Scene Description	Entities (names + layouts) with Consistency Grouping	Background
<b>Scene 1</b>	A cat is lying down on a bed	Frame 1: {'a fluffy Siamese cat': [0.25, 0.25, 1.00, 0.70], 'a plush beige bed': [0.00, 0.50, 1.00, 1.00]} Frame 2: {'a fluffy Siamese cat': [0.25, 0.25, 1.00, 0.70], 'a plush beige bed': [0.00, 0.50, 1.00, 1.00]} ...	Bedroom
<b>Scene 2</b>	Then she gets up	Frame 1: {'a fluffy Siamese cat': [0.55, 0.25, 0.85, 0.55], 'a plush beige bed': [0.00, 0.60, 1.00, 1.00]} Frame 2: {'a fluffy Siamese cat': [0.50, 0.30, 0.80, 0.60], 'a plush beige bed': [0.00, 0.60, 1.00, 1.00]} ...	Bedroom
<b>Scene 3</b>	She goes to the kitchen and eats a snack	Frame 1: {'a fluffy Siamese cat': [0.15, 0.20, 0.40, 0.45], 'gourmet cat snack': [0.50, 0.45, 0.80, 0.65]} Frame 2: {'a fluffy Siamese cat': [0.35, 0.30, 0.60, 0.55], 'gourmet cat snack': [0.50, 0.45, 0.80, 0.65]} ...	Kitchen

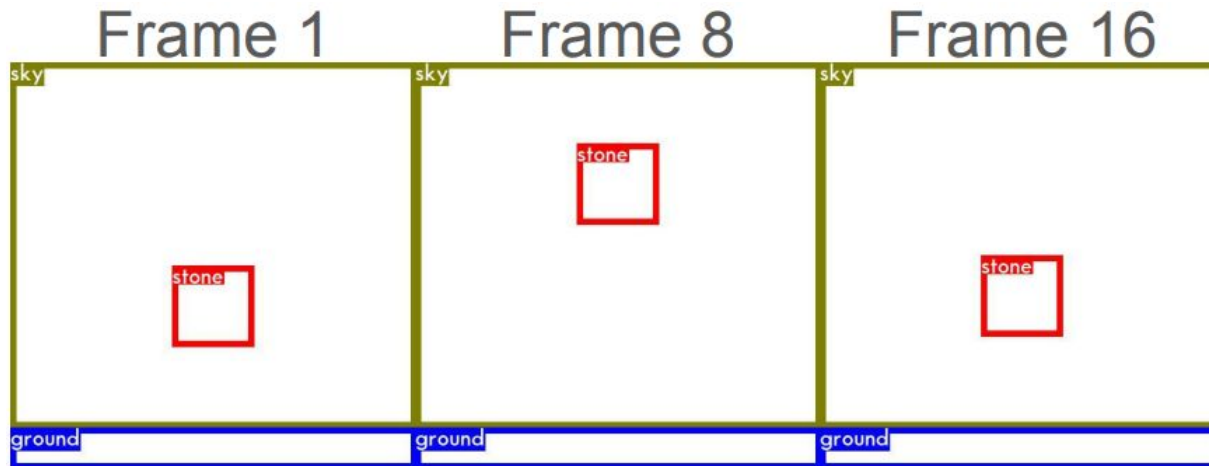
Video Generator



# LLM's Understanding of Basic Physics

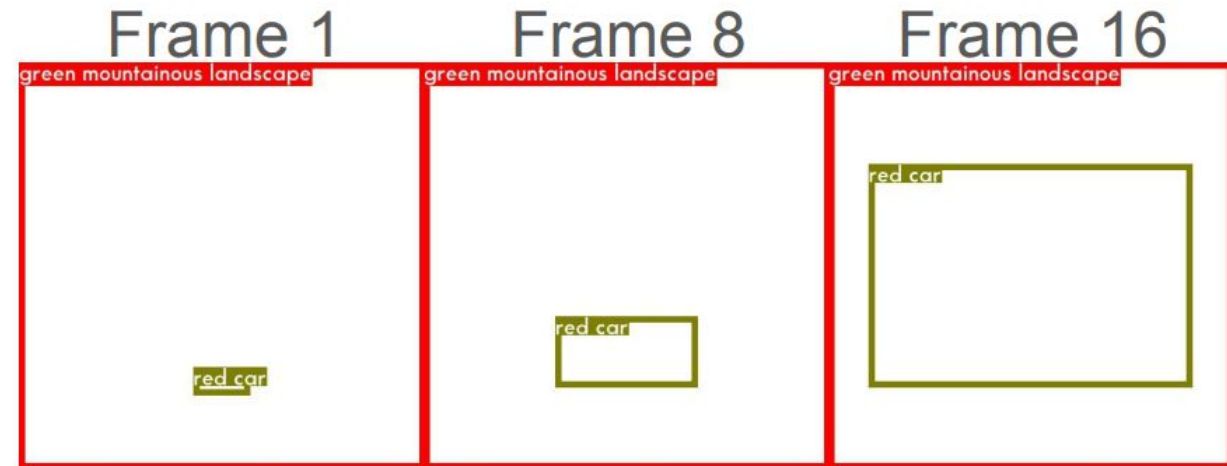
## Gravity

A stone thrown into the sky



## Perspective

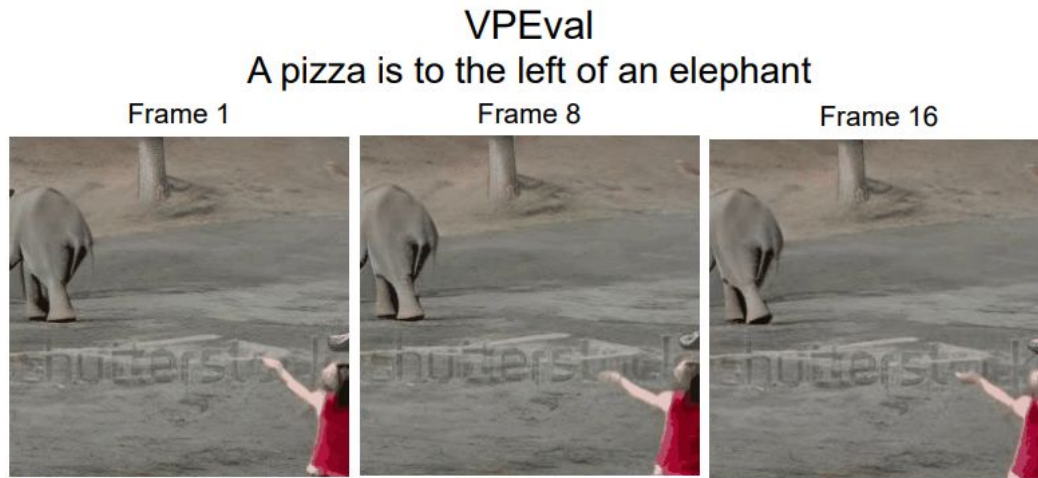
A car is approaching from a distance



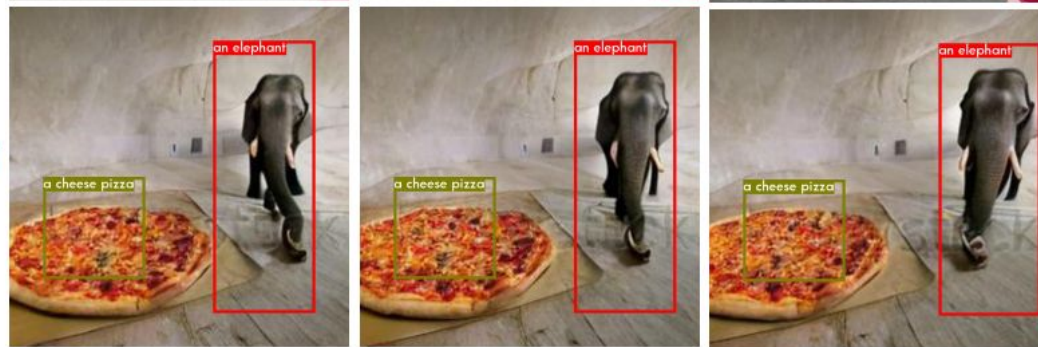


# Object Locations and Movements

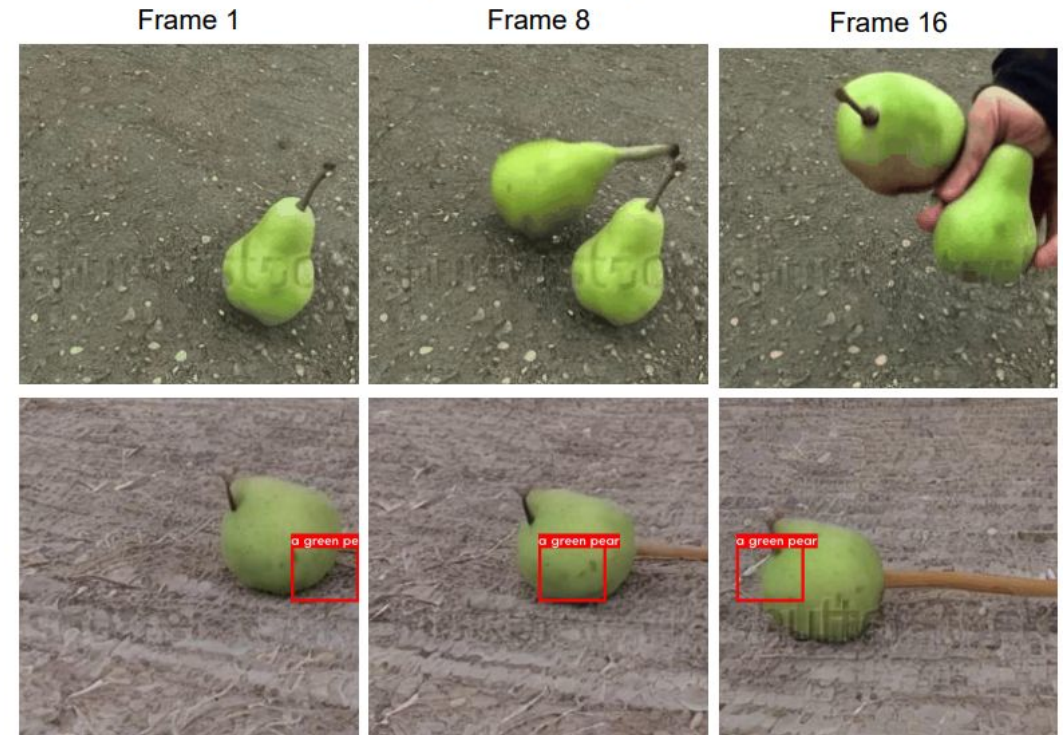
ModelScopeT2V



VideoDirectorGPT  
(Ours)

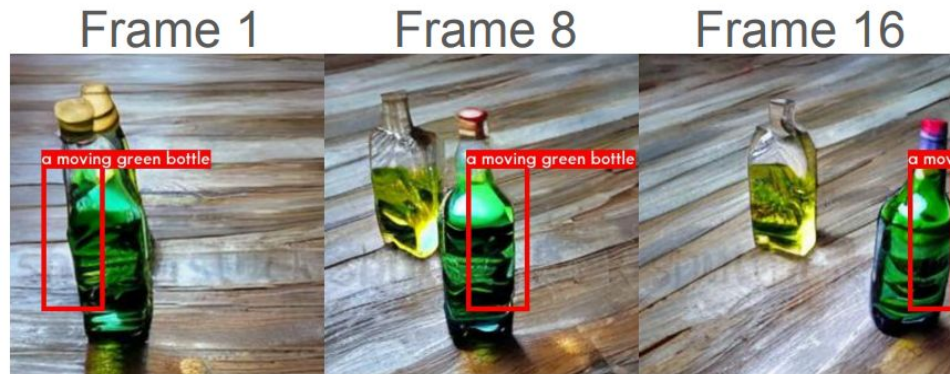


ActionBench-Direction  
Pushing pear from right to left

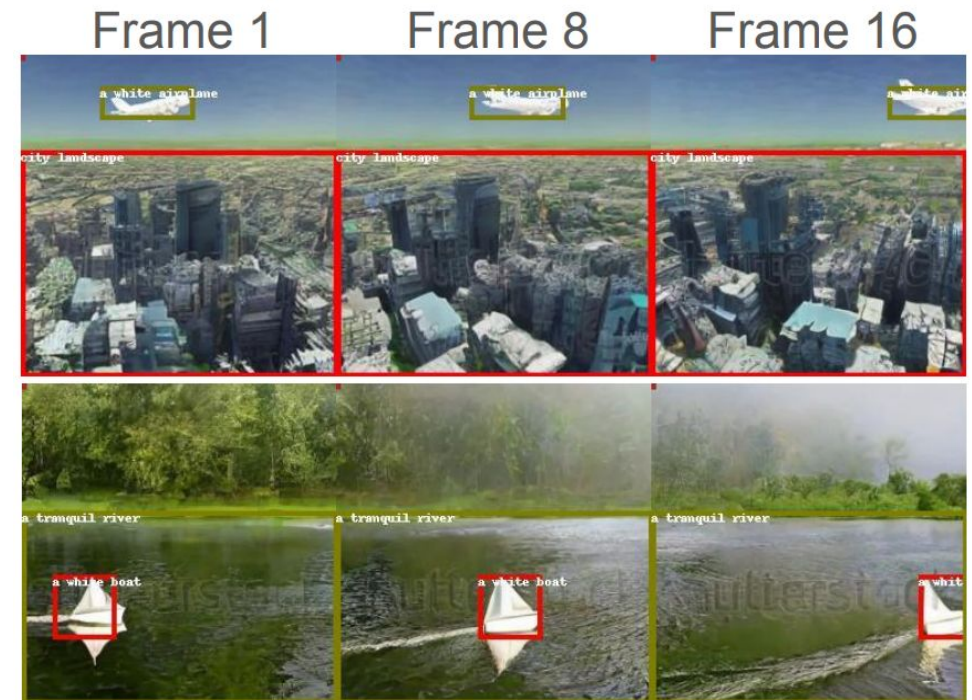


# Movement of Static Objects vs. Objects that Moves

“A {bottle/airplane/boat} moving from left to right.”



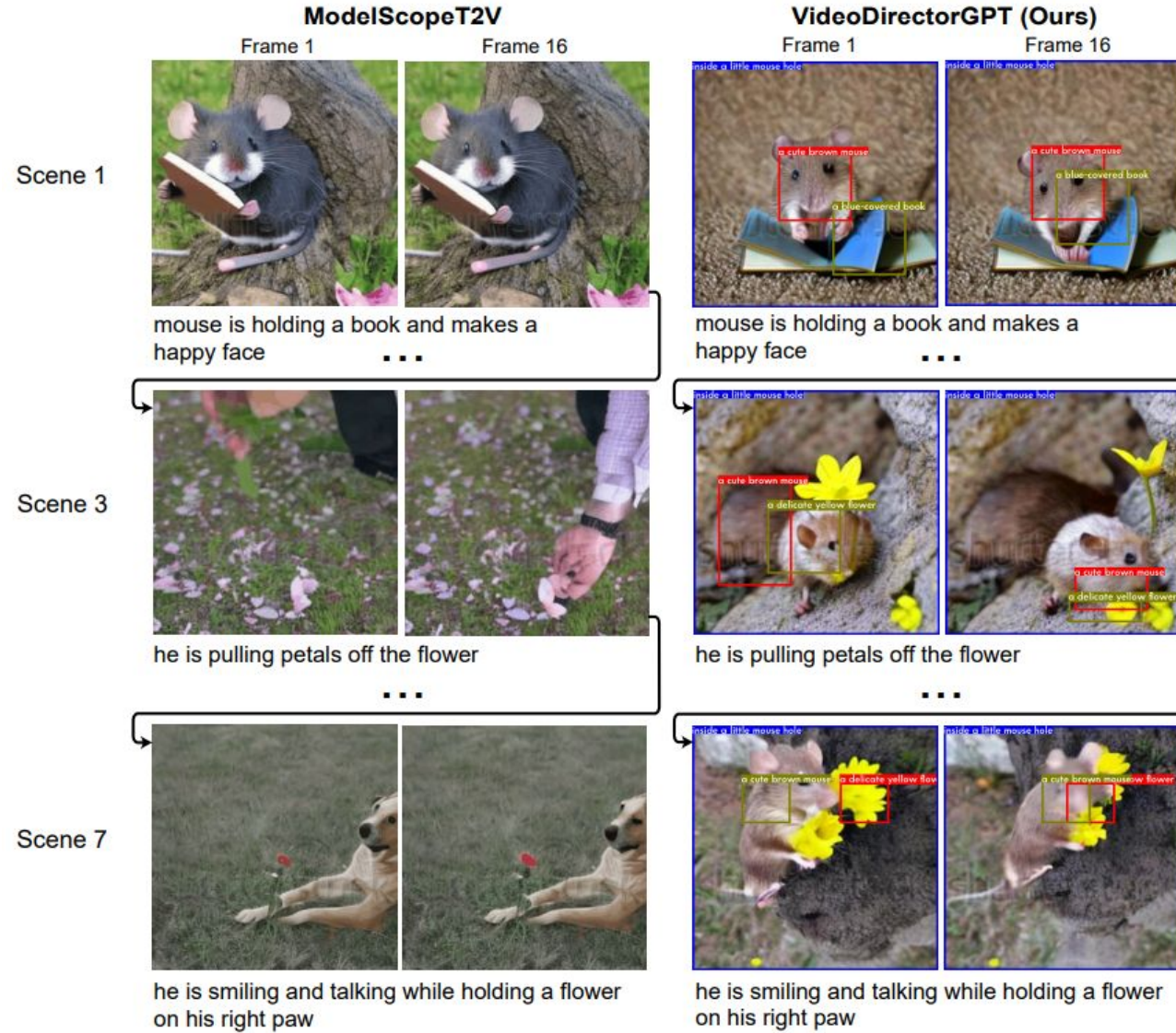
Static objects  
-> Movements of Camera



Objects that can move  
-> Movements of Object (+ Camera)



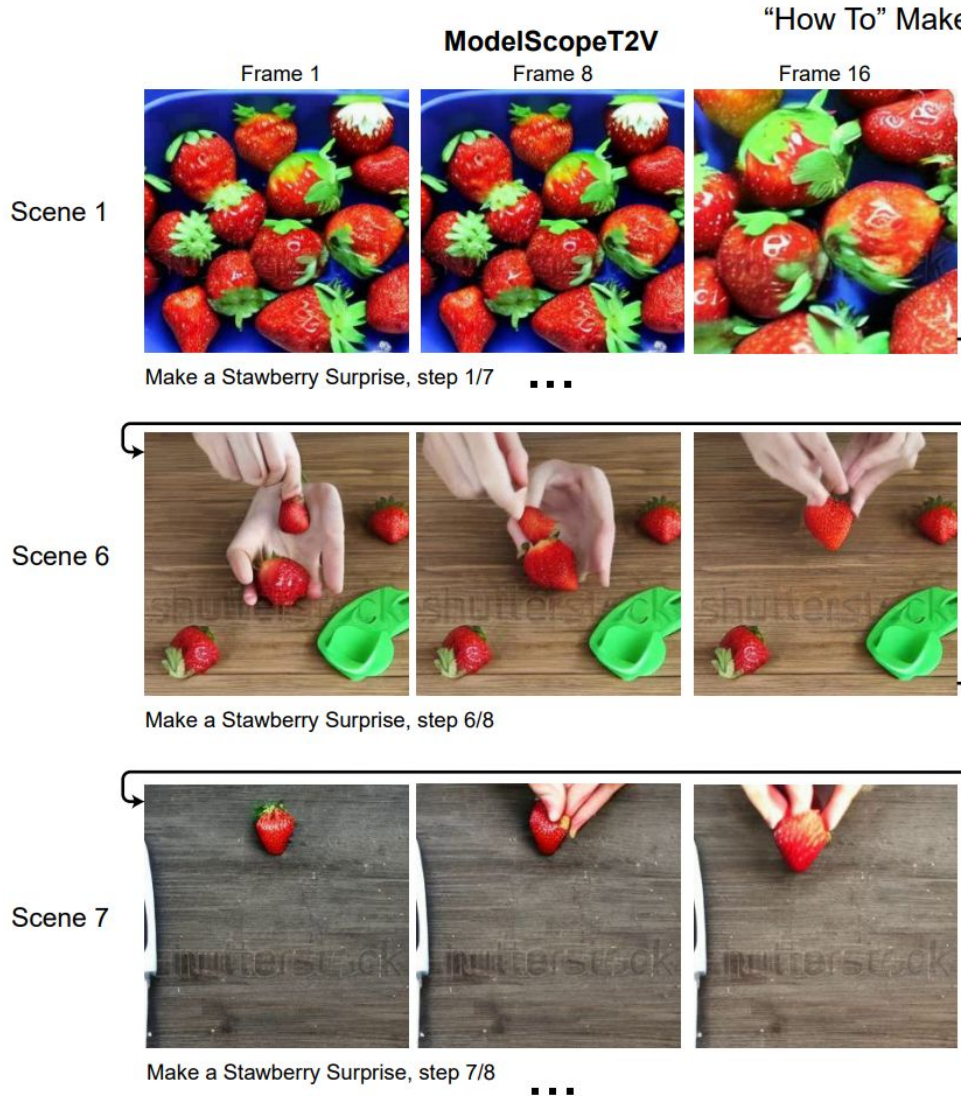
# Multi-Sentence to Multi-Scene Video (Coref-SV)



✗ fails to keep “mouse”  
through all scenes

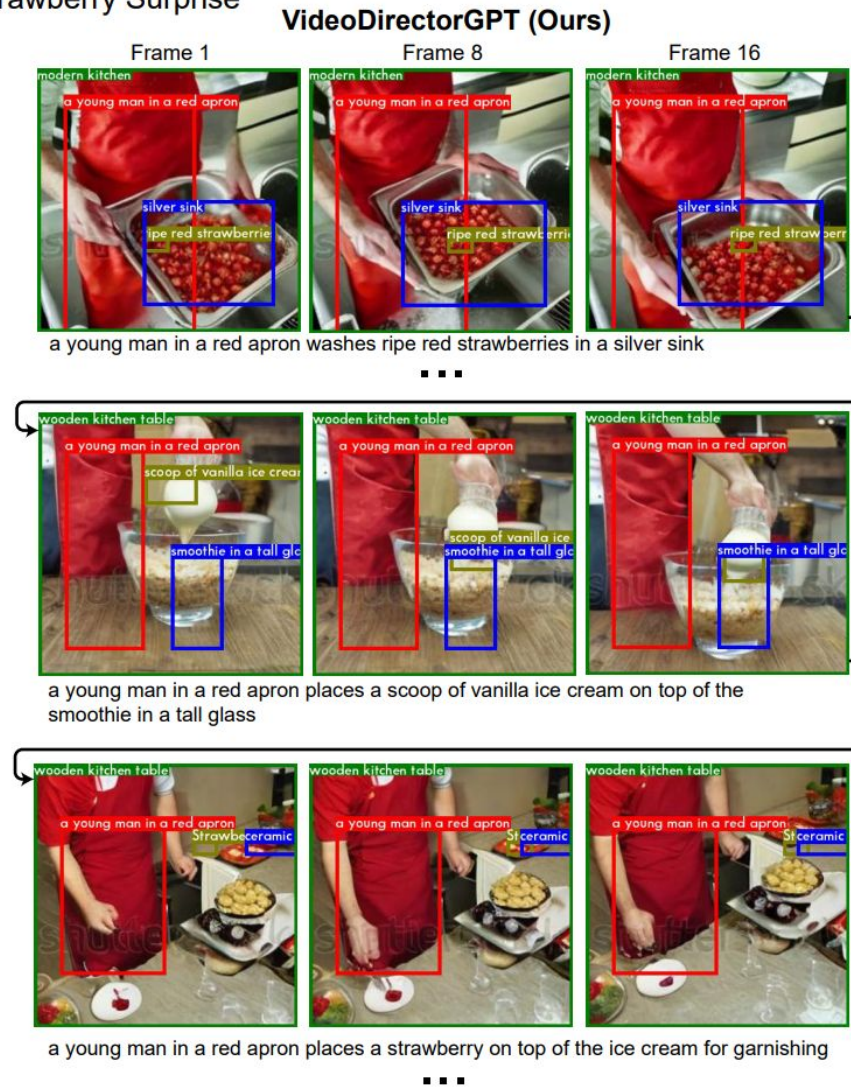
✓ the “mouse” looks consistent  
through all scenes

# Multi-Scene Videos from a Single Sentence



✗ no actual process shown on how to “make” the strawberry surprise dessert

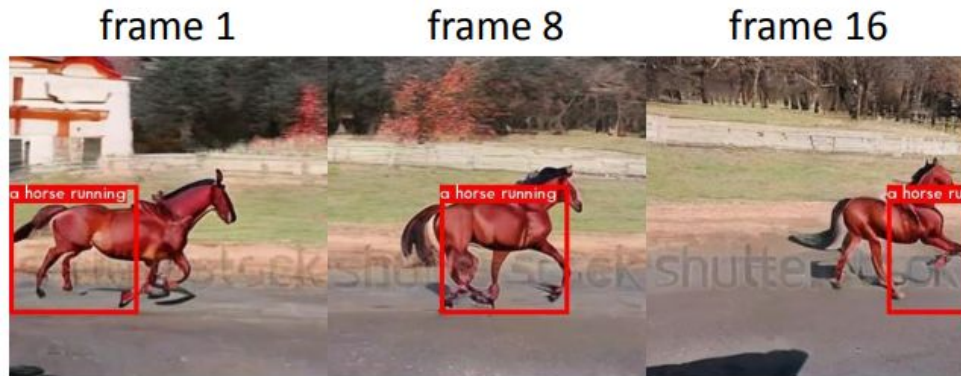
“How To” Make a Strawberry Surprise



✓ step-by-step process on how to “make” the strawberry surprise dessert



# Human-in-the-Loop Video Editing (by modifying video plans)



Original prompt: "A horse running"



Edit 1: Make the horse smaller









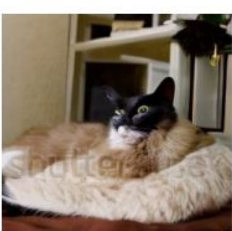
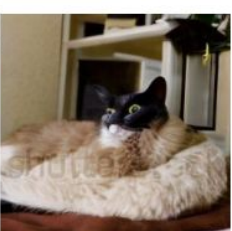





Edit 2: Add "grassland" background



Edit 3: Add "night street" background



# User-Provided Input Image → Video

Entity Grounding		Generated Scenes					
Text Input	<S> = "white cat"	Scene 1: a <S> then gets up from a plush beige bed		Scene 2: a <S> goes to the cream-colored kitchen and eats a can of gourmet cat snack.		Scene 3: a <S> sits next to a large floor-to-ceiling window	
		Frame 1	Frame 16	Frame 1	Frame 16	Frame 1	Frame 16
							
Image+Text Input	 <S> = "cat"						
	 <S> = "cat"						
	 <S> = "teddy bear"						

# Quantitative Evaluation & Human Evaluation

Method	VPEval Skill-based					ActionBench-Direction
	Object	Count	Spatial	Scale	Overall Acc. (%)	Movement Direction Acc. (%)
ModelScopeT2V	89.8	38.8	18.0	15.8	40.8	30.5
VIDEODIRECTORGPT (Ours)	<b>97.1</b>	<b>77.4</b>	<b>61.1</b>	<b>47.0</b>	<b>70.6</b>	<b>46.5</b>

Method	ActivityNet Captions			Coref-SV	HiREST	
	FVD (↓)	FID (↓)	Consistency (↑)	Consistency (↑)	FVD (↓)	FID (↓)
ModelScopeT2V	980	18.12	46.0	16.3	1322	23.79
ModelScopeT2V (with GT co-reference; oracle)	-	-	-	37.9	-	-
VIDEODIRECTORGPT (Ours)	<b>805</b>	<b>16.50</b>	<b>64.8</b>	<b>42.8</b>	<b>733</b>	<b>18.54</b>

Evaluation category	Human Preference (%) ↑		
	VIDEODIRECTORGPT (Ours)	ModelScopeT2V	Tie
Quality	<b>54</b>	34	12
Text-Video Alignment	<b>54</b>	28	18
Object Consistency	<b>58</b>	30	12

# VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning

[videodirectorgpt.github.io](https://videodirectorgpt.github.io)

**Han Lin**



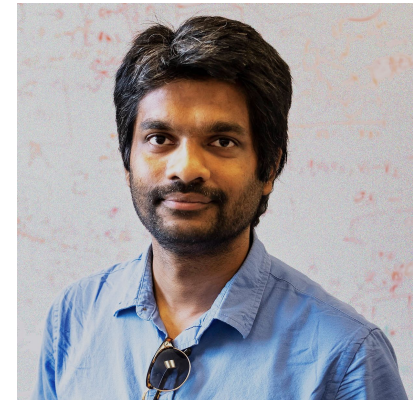
**Abhay Zala**



**Jaemin Cho**



**Mohit Bansal**



THE UNIVERSITY  
of NORTH CAROLINA  
at CHAPEL HILL