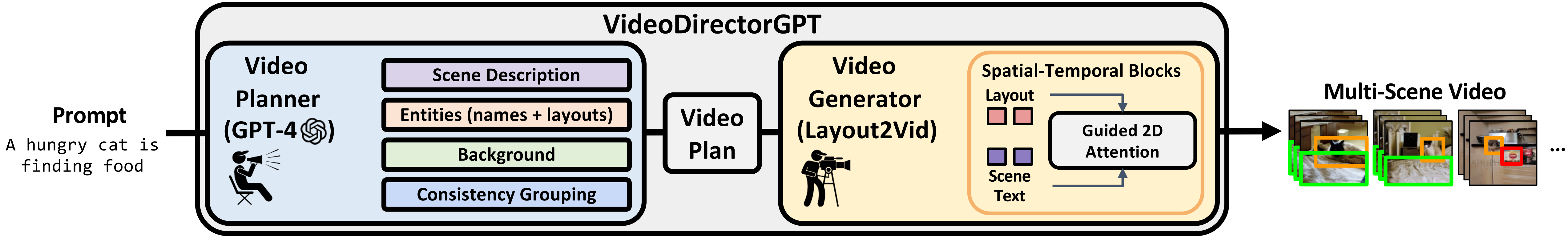
VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning



https://videodirectorgpt.github.io/

[Lin et al., 2023]

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

Prompt

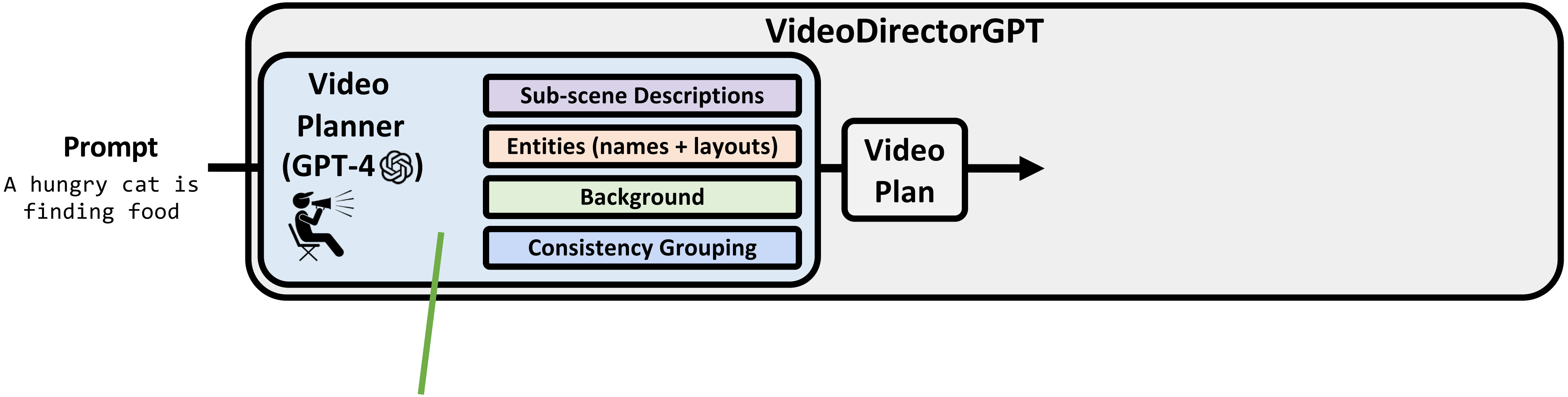
A hungry cat is finding food

Single input text prompt

https://videodirectorgpt.github.io/

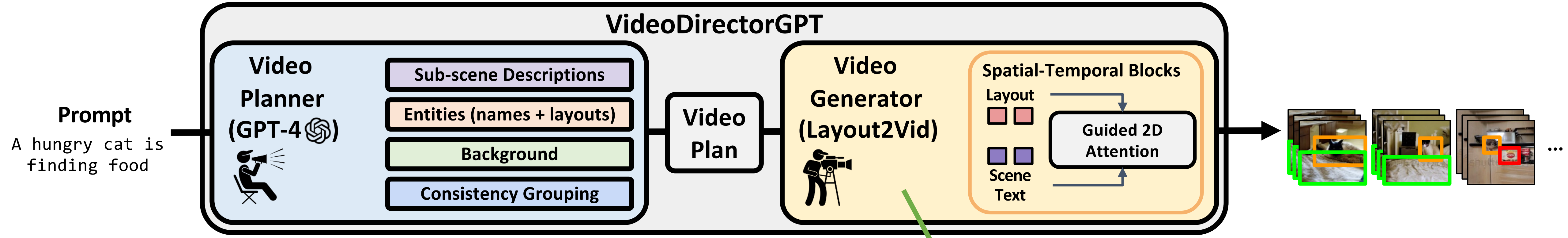
[Lin et al., 2023]

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.

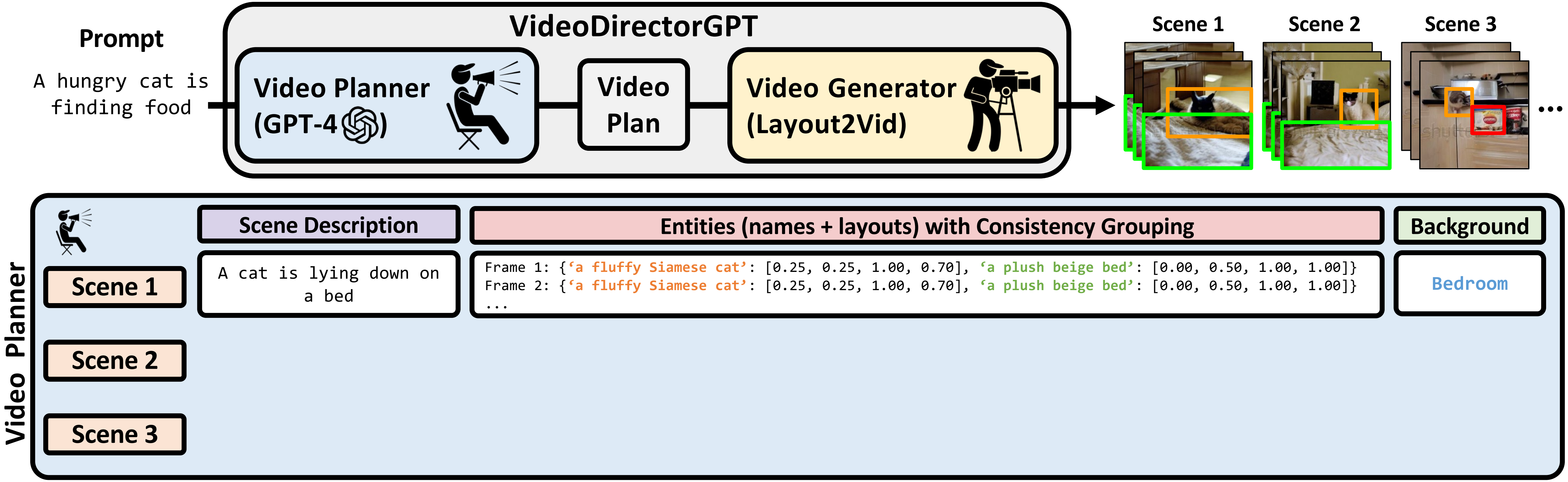
VideoDirectorGPT: Consistent Multi-Scene Video Generation via LLM-Guided Planning



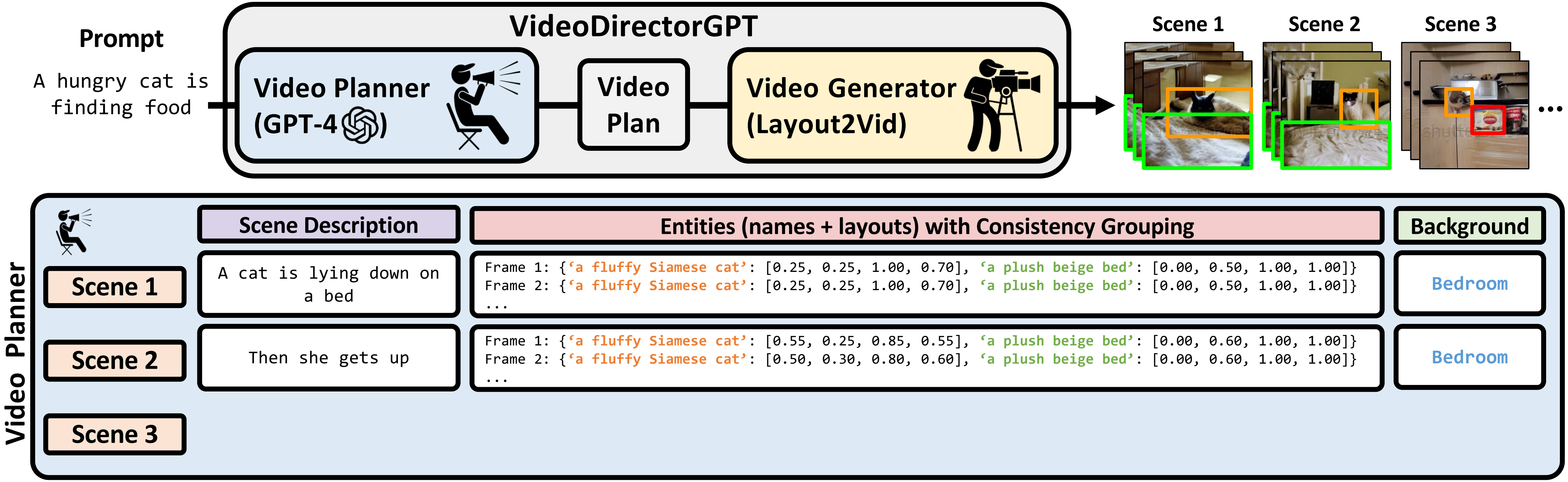
- Takes the ***video plan***
- Generates the video
  - Follows the 2D bbox layouts
  - Maintains visual consistency

https://videodirectorgpt.github.io/

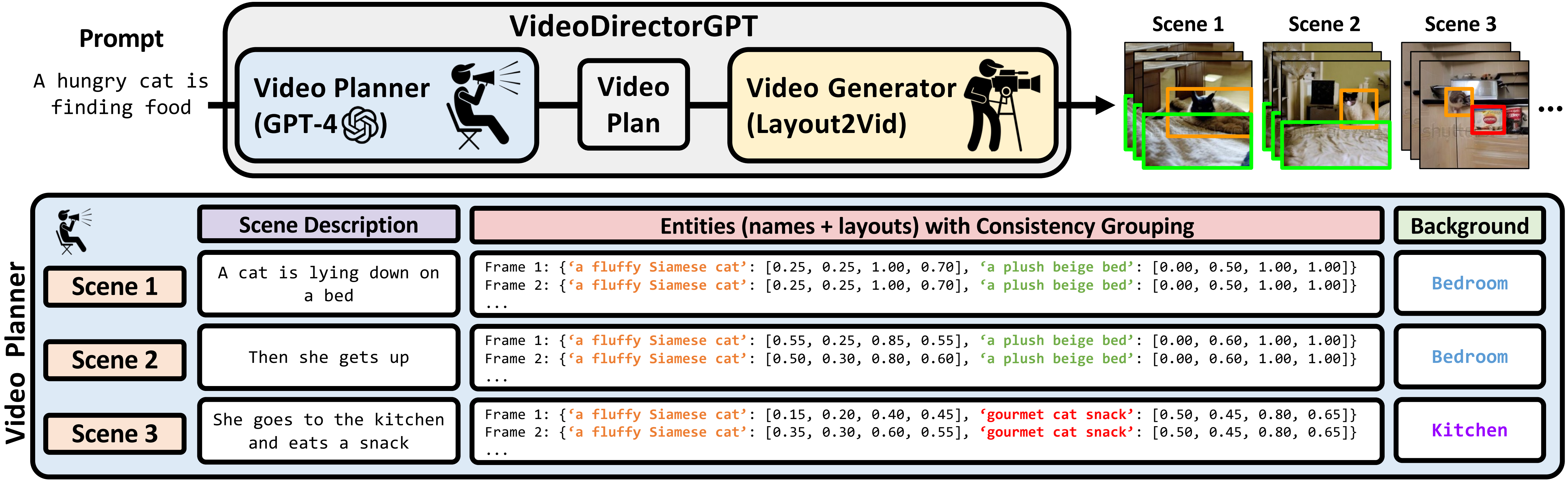
[Lin et al., 2023]

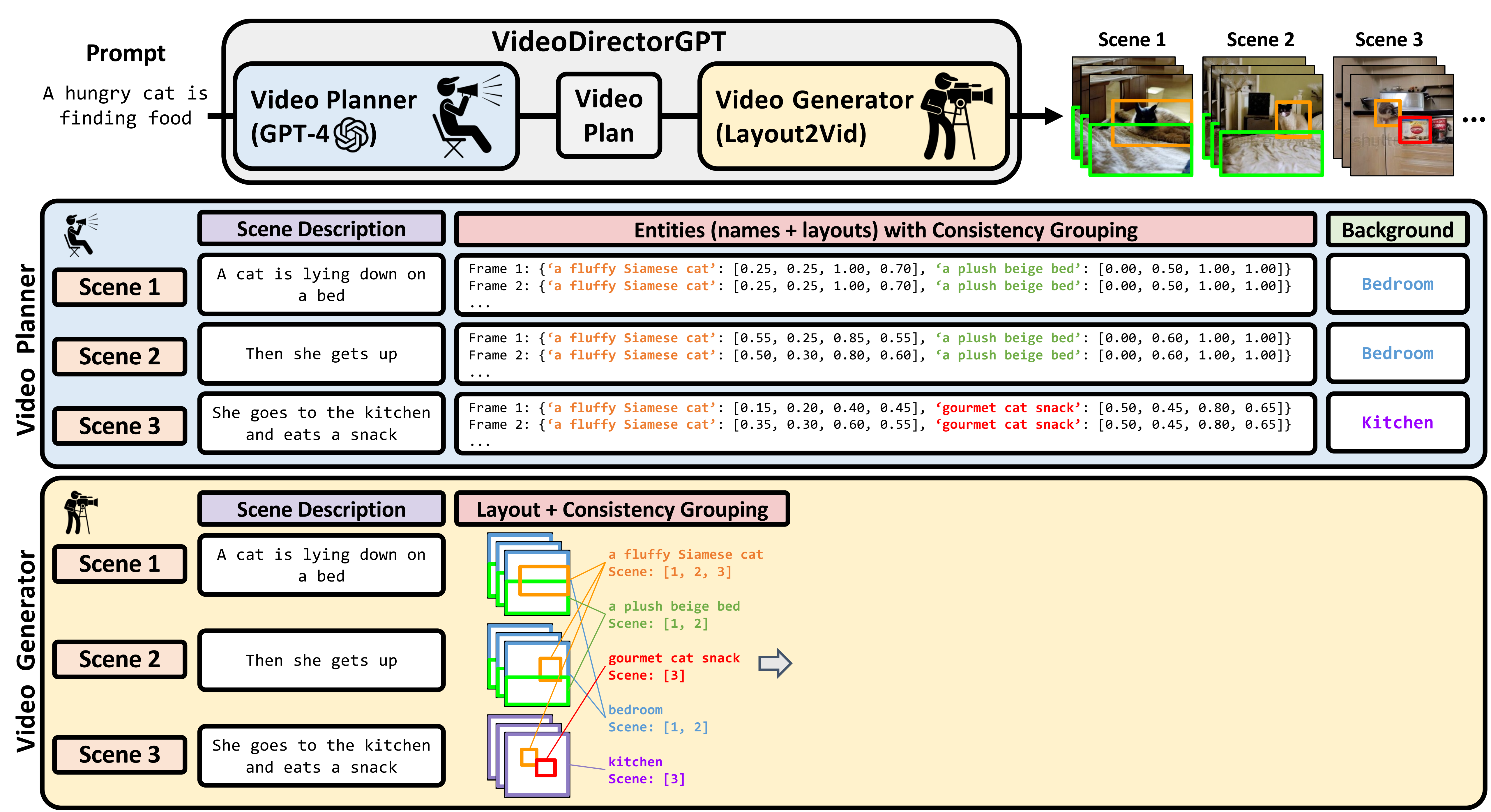


https://videodirectorgpt.github.io/ [Lin et al., 2023]

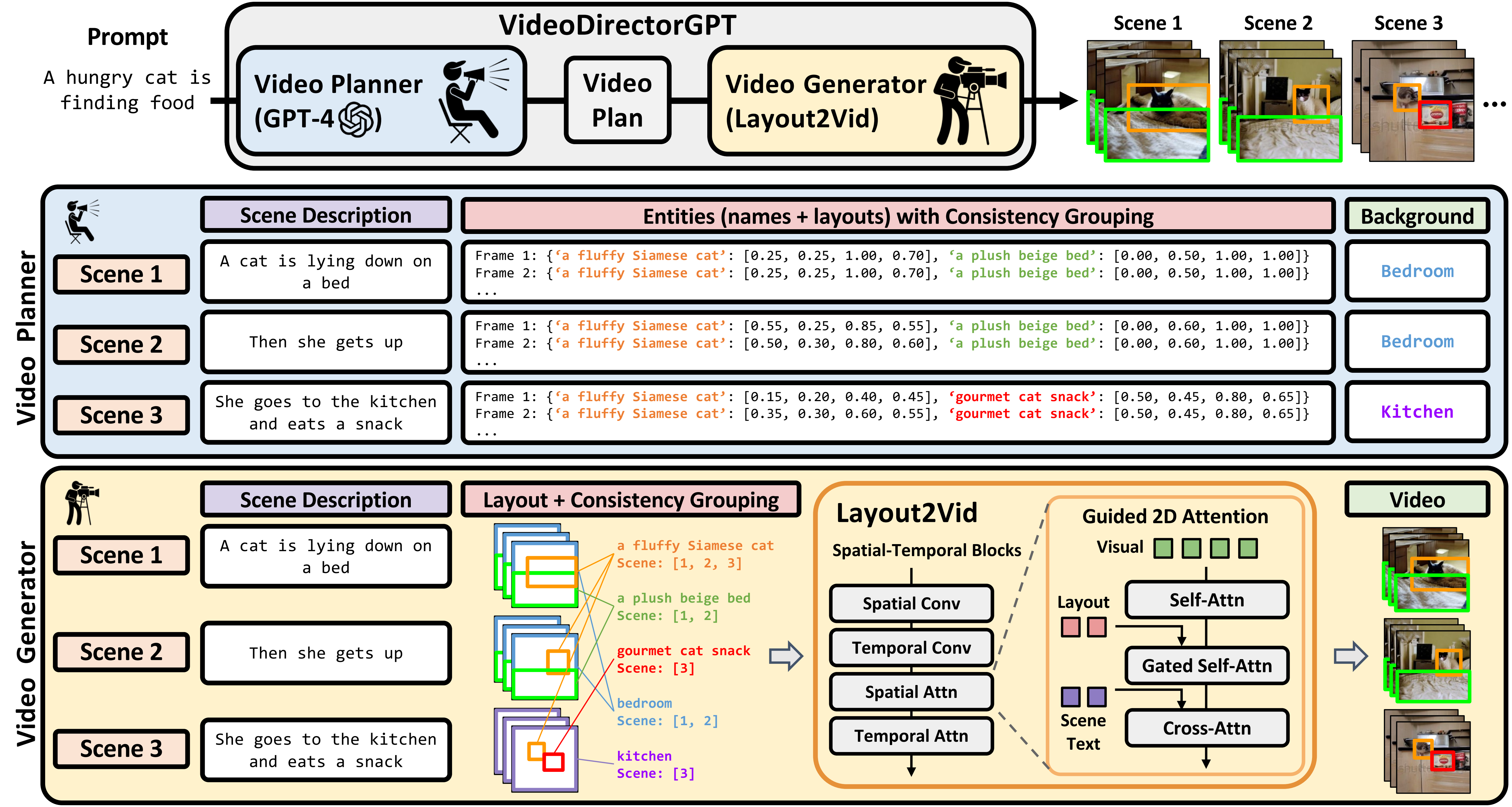


https://videodirectorgpt.github.io/ [Lin et al., 2023]


Prompt
A hungry cat is finding food
VideoDirectorGPT
Video Planner (GPT-4)
Video Plan
Video Generator (Layout2Vid)
Scene 1
Scene 2
Scene 3
...
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



https://videodirectorgpt.github.io/ [Lin et al., 2023]



https://videodirectorgpt.github.io/ [Lin et al., 2023]

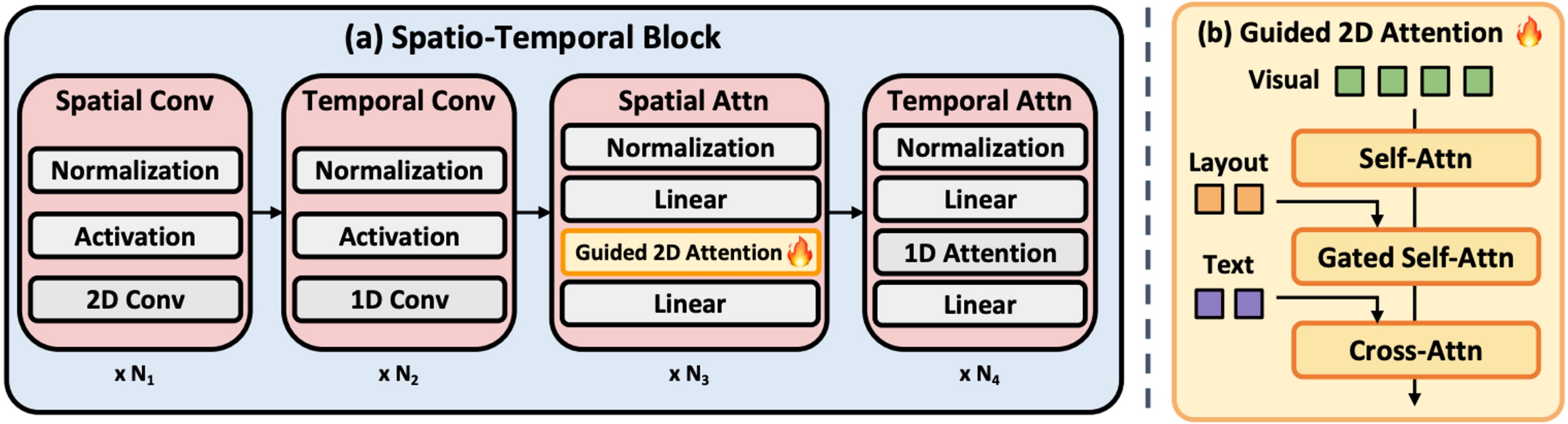
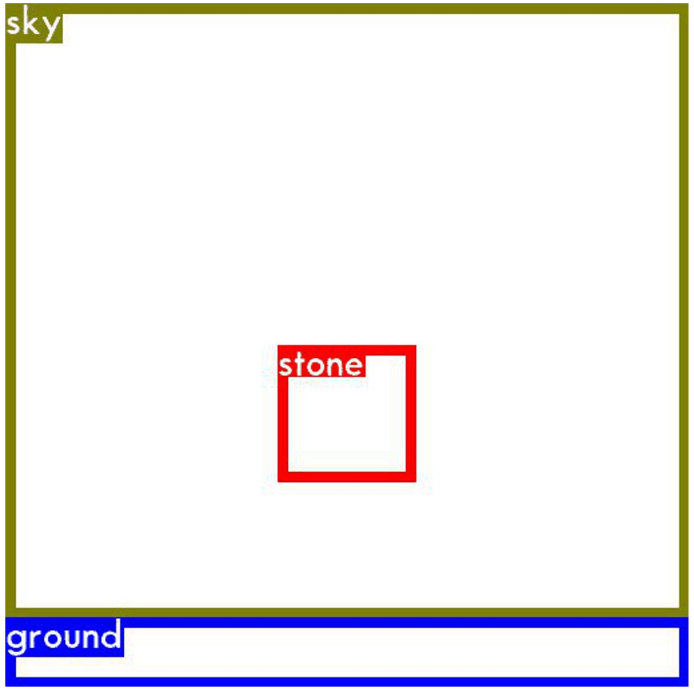


Figure 3: Overview of **(a) spatio-temporal blocks** within the diffusion UNet of our **Layout2Vid** and **(b) Guided 2D Attention** present in the spatial attention module. (a) The spatio-temporal block comprises four modules: spatial convolution, temporal convolution, spatial attention, and temporal attention. We adopt settings from ModelScopeT2V, where (N1, N2, N3, N4) are set to (2, 4, 2, 2). In (b) Guided 2D Attention, we modulate the visual representation with layout tokens and text tokens. For efficient memory usage and training, only the parameters of the Guided 2D Attention (indicated by the fire symbol, constituting 13% of total parameters) are trained using image-level annotations. The remaining modules in the spatio-temporal block are kept frozen.

LLM’s Understanding of Basic Physics

Gravity

A stone thrown into the sky



Perspective

A car is approaching from a distance

https://videodirectorgpt.github.io/

[Lin et al., 2023]

Object Movement

a **pear** moving from **right to left**

ModelScopeT2V

❌ fails to move the "pear"

VideoDirectorGPT (Ours)

✓ correctly moves the the "pear" from right to left

https://videodirectorgpt.github.io/ [Lin et al., 2023]

Movement of Static Objects vs. Objects that Moves

"A {**bottle/airplane**} moving from **left to right**."

Static objects
-> Movements of Camera

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

https://videodirectorgpt.github.io/

[Lin et al., 2023]

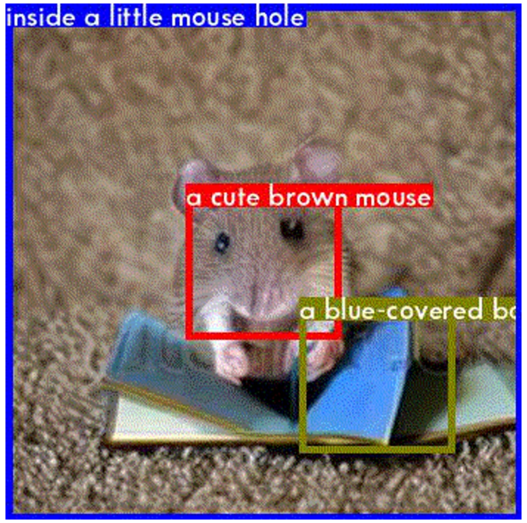
Multi-Sentence to Multi-Scene Video (Coref-SV)

Scene 1: **mouse** is holding a book and makes a happy face.
Scene 2: **he** looks happy and talks.
Scene 3: **he** is pulling petals off the flower.
Scene 4: **he** is ripping a petal from the flower.
Scene 5: **he** is holding a flower by **his** right paw.
Scene 6: one paw pulls the last petal off the flower.
Scene 7: **he** is smiling and talking while holding a flower on **his** right paw.

ModelScopeT2V

❌ fails to keep "mouse" through all scenes

VideoDirectorGPT (Ours)



✓ the "mouse" looks consistent through all scenes

Multi-Scene Videos from a Single Sentence

make a strawberry surprise

Video Planner (GPT-4)

Generated multi-scene prompts (total 10 scenes):

1. A bartender prepares the working area by cleaning and organizing.
2. The bartender rinses fresh strawberries under a tap.
3. The bartender cuts the strawberries and removes the stems.

…

Video Generator (Layout2Vid)

ModelScopeT2V (baseline)

❌ no actual process shown on how to "make" the strawberry surprise dessert

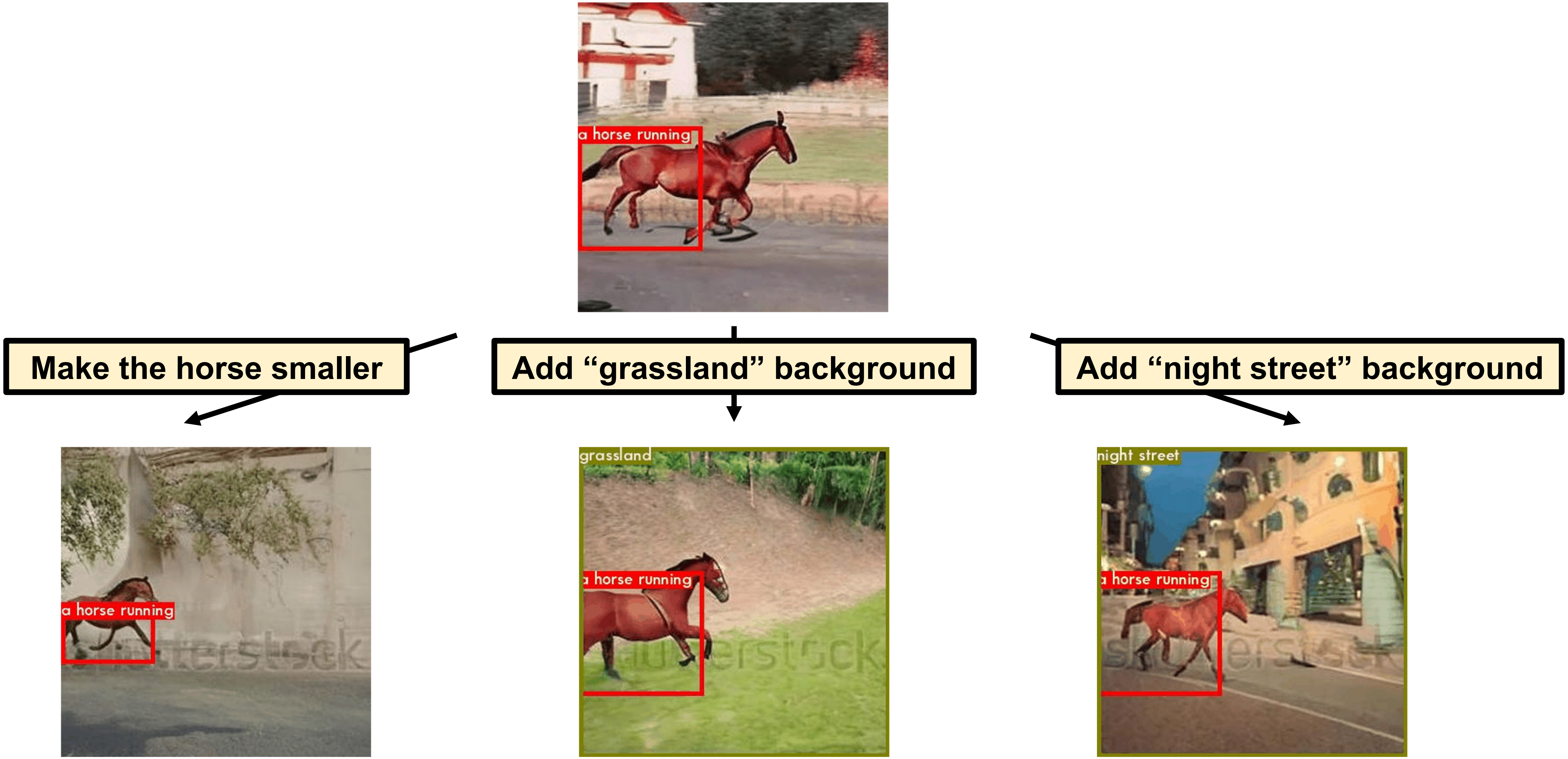
VideoDirectorGPT (Ours)

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

https://videodirectorgpt.github.io/

[Lin et al., 2023]

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



https://videodirectorgpt.github.io/

[Lin et al., 2023]

User-Provided Input Image ⟶ Video

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 snack.
Scene 3: a <S> sits next to a large floor-to-ceiling window.

<S> = “cat”
+

<S> = “teddy bear”
+

https://videodirectorgpt.github.io/

[Lin et al., 2023]

Quantitative 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	**97.1**	**77.4**	**61.1**	**47.0**	**70.6**	**46.5**

Object movement direction accuracy:

- First obtain the start/end locations of objects via GroundingDINO on the first/last video frames
- Then evaluate whether the x and y coordinates of the objects have changed correctly as described in the prompts (through a binary score of 0 or 1)

Quantitative Evaluation

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)	**805**	**16.50**	**64.8**	**42.8**	**733**	**18.54**

Multi-scene object consistency:

- First detect the target object from the center frame of each scene
- Then extract the CLIP image embedding from the detected bounding box
- Calculate the consistency metric by averaging the CLIP image embedding similarities across all adjacent scene pairs

$$\frac{1}{N}\sum_{n=1}^{N-1} cos(\mathrm{CLIP}_n^{\mathrm{img}}, \mathrm{CLIP}_{n+1}^{\mathrm{img}})$$

https://videodirectorgpt.github.io/

[Lin et al., 2023]

Human Evaluation

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

https://videodirectorgpt.github.io/

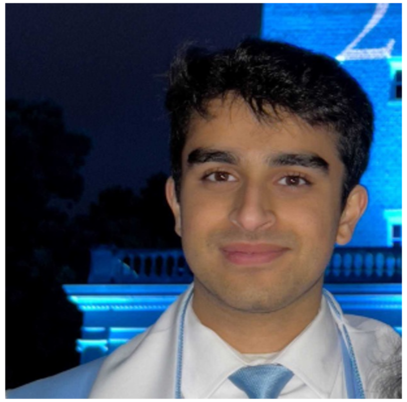
[Lin et al., 2023]

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

videodirectorgpt.github.io

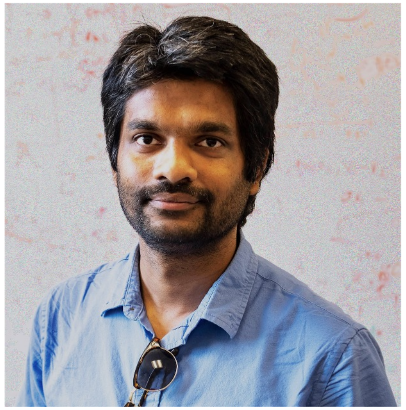
Han Lin

Abhay Zala



Jaemin Cho

Mohit Bansal



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL