VideoGLaMM: A Large Multimodal Model for Pixel-Level Visual Grounding in Videos

Supplementary Material

We provide supplementary material for further understanding of certain sections of the main paper. We have arranged the sections as follows:

- Additional Quantitative Results
- Additional Qualitative Results
- Instruction templates for Annotation pipeline
- Ethics and Societal Impact

A. Additional Quantitative Results

We perform additional quantitative evaluations of VideoGLaMM on the Refer-YouTube-VOS dataset for the task of referring video object segmentation. As discussed in the Sec. 5.2 of the main paper, for referring video segmentation, the output should be grounded as per the given phrase, pointing towards specific instances in the video. Table 8 shows the results on Refer-Youtube-VOS dataset. Our VideoGLaMM outperforms all the contemporary best performing baselines, suggesting the enhanced grounding and localization capability of our model.

Model	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
LISA-7B [20]	50.6	49.7	50.2
LISA-13B [20]	53.0	52.1	52.6
TrackGPT-7B [58]	57.4	55.3	56.4
TrackGPT-13B [58]	60.8	58.1	59.5
VideoLISA [5]	65.7	61.7	63.7
VideoGLaMM	**65.4**	**68.2**	**66.8**

Table 8. **Performance comparison of VideoGLaMM on Refer-Youtube-VOS:** VideoGLaMM shows superior performance on segmenting referring objects in the videos.

B. Additional Qualitative Results

B.1. Qualitative results on GCG and referring object segmentation

Figure 6 and Figure 7 shows our model's performance on the GCG and referring object segmentation tasks respectively.

B.2. Conditional Video Generation

We incorporate VideoGLaMM into a conditional video generation model [45], designed to take an input video and modify it based on a specified condition (in this case, a mask) and a query prompt that outlines the desired edits. Given a video and a plain text query, VideoGLaMM identifies the relevant objects in the video by generating a precise mask for the object of interest. The generative model then utilizes the video, mask, and query prompt to produce the edited video. Figure 5 illustrates two examples: object removal and object replacement, with the masks provided by VideoGLaMM. This demonstrates VideoGLaMM's adaptability in working with different models to address a wide range of video editing tasks.

C. Instruction Prompt templates for Annotation pipeline

We provide the instruction prompt templates used at each stage our semi-automatic pipeline below.

a) Videos with only masks.

i) Prompt for generating object patch descriptions

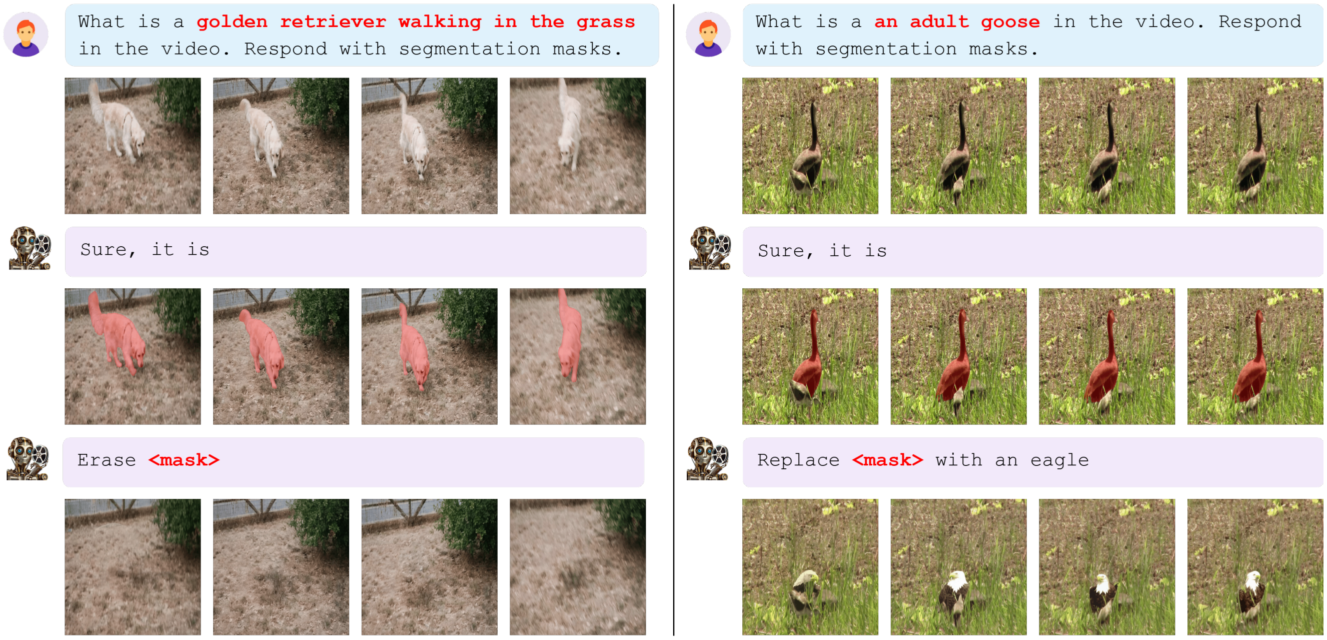
`These are frames from a video that I want to upload. What does the {class_name} look like, and what is the {class_name} doing?`

ii) Prompt for generating refined object descriptions

`These are frames from a video that I want to upload. Please refine this caption: {caption from step 1}. The instance in the video is highlighted by a rectangular box with the color corresponding to ID {object_id}`

iii) Prompt for generating caption

`These are frames from a video that I want to upload. In the video, the ID number of the box is on the top left of the box. There are some instance captions:[The obj_{object_id} must be surrounded by a rectangular box with color number {object_id}. It is a {class_name}.{object_id's caption from step 2}, The obj_{object_id} must be surrounded by a rectangular box with color number {object_id}. It is a {class_name}.{object_id's caption from step 2}...] Generate a dense caption that describes the video in detail based on the video and instance captions, including all of`



What is a **golden retriever walking in the grass** in the video. Respond with segmentation masks.

Sure, it is

Erase **<mask>**

What is a **an adult goose** in the video. Respond with segmentation masks.

Sure, it is

Replace **<mask>** with an eagle

a) Object Erasing

a) Object Replacement

Figure 5. **Conditional Video Generation using VideoGLaMM**.

the instances mentioned in the instance captions and other instances in the video. Ensure that each instance mentioned in the instance caption appears exactly once in the dense caption, followed by the format {obj_} to indicate which instance caption the mentioned instance corresponds to. The {obj_} must directly follow the noun representing the instance

---

b) Videos having Bbox annotations and captions

Your task is to process video captions to make them more detailed and explanatory. You are given a ground truth caption and a set of dense noisy captions. Ground truth caption contains a description of the objects visible in a video, with noun phrases of significant objects surrounded by <p> and </p> tags, followed by a [SEG:x] tag.
Dense noisy captions contain additional information about the video, but they may be redundant or less precise than the ground truth caption.
Your task is to paraphrase the ground truth caption by incorporating relevant information from the dense noisy captions.
The refined caption should be more detailed and explanatory than the ground truth caption.
The refined caption should preserve the original <p>, </p>, and [SEG:x] tags.
The refined caption should also preserve the identity of [SEG:x] tags, given by a unique identifier x.

You may look at the following examples:
Example 1:
Ground truth caption:
A <p> weight </p> [SEG:1] lifter is in a <p> gym </p> [SEG:2] , and <p> he </p> [SEG:1] lifts a <p> barbell </p> [SEG:0]
Reference captions:
In the video, a man is lifting weights in a gym. He lifts the weights over his head and then drops them on the ground.
In the video, a person is seen lifting weights in a gym setting. The individual is focused on performing the weightlifting exercise, and their posture indicates a controlled and deliberate movement. The gym environment is equipped with various weightlifting equipment, and there are other people present in the background, suggesting a shared workout space. The person's attire and the equipment indicate that this is a dedicated space for physical fitness and strength training. The video captures a moment of physical exertion and dedication to fitness.
Output:

{"refined_caption": "In the video, <p> A man </p> [SEG:1] is lifting weights in a <p> gym </p> [SEG:2]. <p> He </p> [SEG:1] is lifting a <p> barbell </p> [SEG:0] over his head and then drops them on the ground."}
Example 2:
Ground truth caption:
The <p> man </p> [SEG:1] stands while holding onto the <p> swing </p> [SEG:0]
Reference captions:
In the video, a man is swinging on a swing set in a park. He is wearing a black shirt and is swinging back and forth while looking towards the camera.
In the video, a person is standing in a park, wearing a black shirt and dark pants. The individual appears to be posing or standing still, possibly enjoying the surroundings or waiting for someone. The park features a playground with visible equipment, such as a swing set, indicating a recreational area for children and families. The person is standing on a concrete surface, and there are trees and other greenery in the background, suggesting a peaceful and natural setting. The individual's pose and the environment create a calm and leisurely atmosphere.
Output:
{"refined_caption": "In the video, <p> a man </p> [SEG:1] is swinging on a <p> swing set </p> [SEG:0] in a park. He is wearing a black shirt and is swinging back and forth while looking towards the camera."}
Example 3:
Ground truth caption:
<p> She </p> [SEG:1] puts shaving <p> cream </p> [SEG:2] on <p> her </p> [SEG:1] <p> leg </p> [SEG:0] and shaves <p> her </p> [SEG:1] <p> leg </p> [SEG:0]
Reference captions:
In the video, a person is seen sitting on a tub and shaving their legs with a razor.
In the video, a person is seen sitting in a bathtub, and their legs are being shaved with a razor. The individual appears to be focused on the shaving process, and there are no other significant actions or events occurring in the video. The person's posture and the position of the razor suggest a careful and deliberate approach to shaving their legs. The setting appears to be a private bathroom, and there are no other people or objects visible in the frame.
Output:
{"refined_caption": "In the video, <p> a woman </p> [SEG:1] is seen sitting in a bathtub, shaving <p> her </p> [SEG:1] <p> legs </p> [SEG:0] with a razor. <p> She </p> [SEG:1] is applying <p> shaving cream </p> [SEG:2] on <p> her </p> [SEG:1] <p> leg </p> [SEG:0]."}

Now please refine the following caption:
Ground truth caption:
{gt_caption}
Reference captions:
{reference_captions}
Please provide the refined caption in (JSON format, with a key refined_caption.)

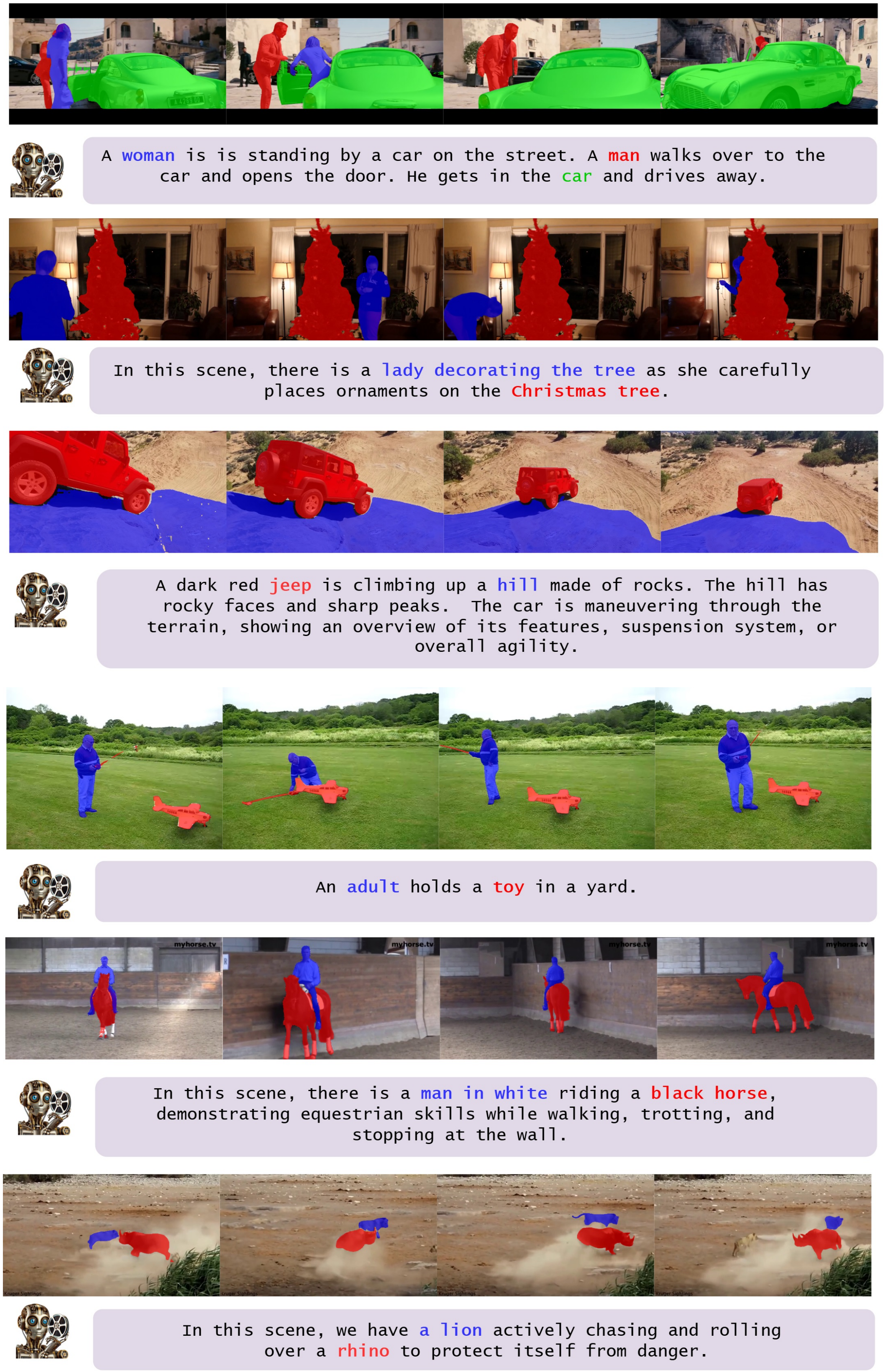
---

c) Videos having Bbox annotations and referring expressions

Your task is to generate annotated video captions, given original unannotated video descriptions, the lists of subjects/objects in the video and the relation between them.
For each video, you are given a relation between a subject and an object, along with the categories and target IDs of the subject and object. Your task is to generate a new caption annotating the subject and object in the caption with the corresponding target IDs.

You may look at the following examples:
Example 1:
Input :
subject :
target_id :0, category : rabbit
object :
target_id : 1, category : adult
relation : lean_on
description: "there is a white rabbit leaning on an adult by the water".
Output:
'caption': 'there is a [white rabbit](0) leaning on an [adult](1) by the water'

Now please process the following.
video_relation_data
In the new caption, the noun phrases should be included within square brackets and object ID/IDs should be included within paranthesis. E.g. [noun phrase](object ID/IDs) .
Please provide the generated caption in JSON format, with a key "caption".



A woman is is standing by a car on the street. A man walks over to the car and opens the door. He gets in the car and drives away.

In this scene, there is a lady decorating the tree as she carefully places ornaments on the Christmas tree.

A dark red jeep is climbing up a hill made of rocks. The hill has rocky faces and sharp peaks. The car is maneuvering through the terrain, showing an overview of its features, suspension system, or overall agility.

An adult holds a toy in a yard.

In this scene, there is a man in white riding a black horse, demonstrating equestrian skills while walking, trotting, and stopping at the wall.

In this scene, we have a lion actively chasing and rolling over a rhino to protect itself from danger.

Figure 6. **Qualitative results of VideoGLaMM on GCG samples**.

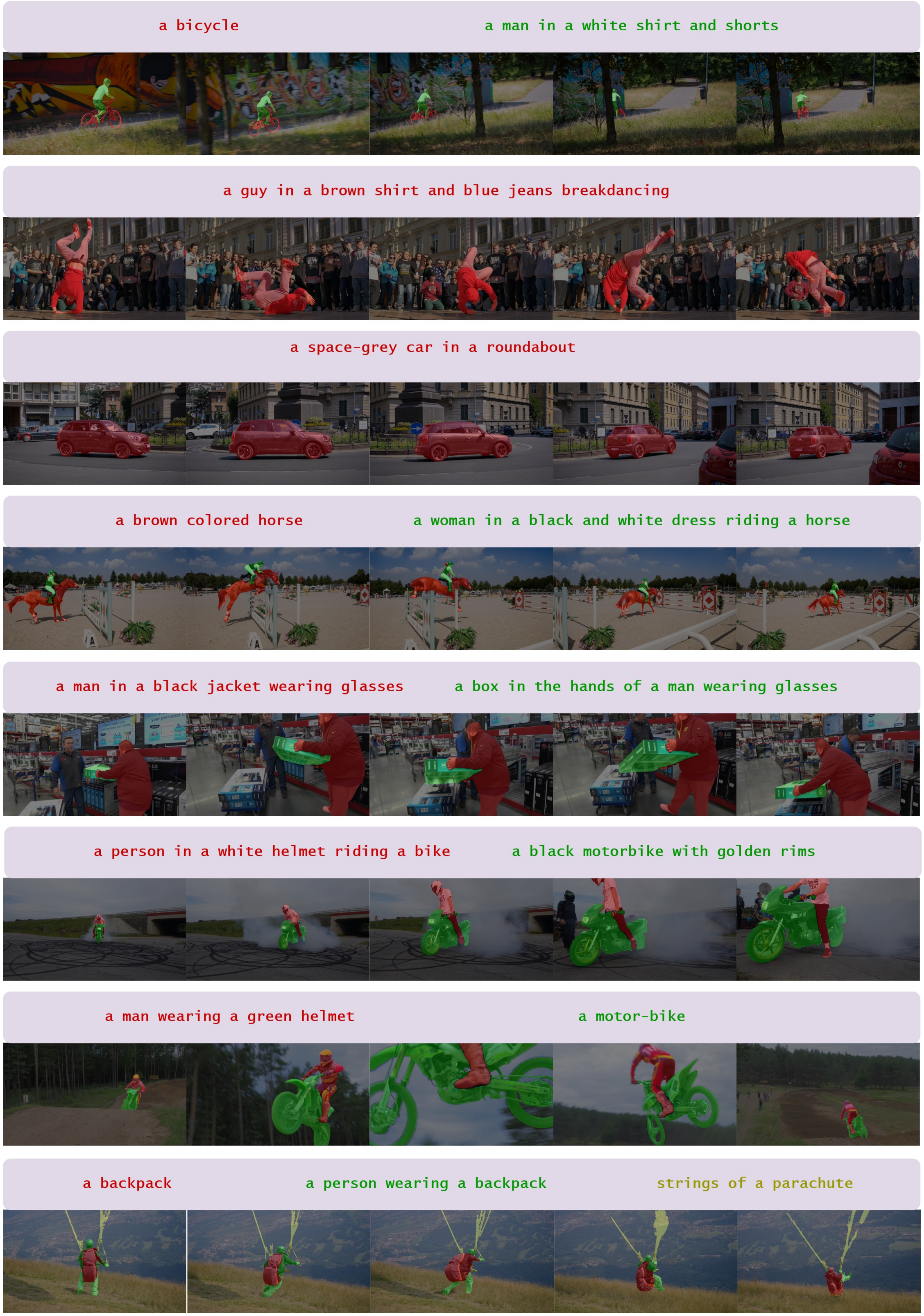

a bicycle
a man in a white shirt and shorts
a guy in a brown shirt and blue jeans breakdancing
a space-grey car in a roundabout
a brown colored horse
a woman in a black and white dress riding a horse
a man in a black jacket wearing glasses
a box in the hands of a man wearing glasses
a person in a white helmet riding a bike
a black motorbike with golden rims
a man wearing a green helmet
a motor-bike
a backpack
a person wearing a backpack
strings of a parachute

Figure 7. **Qualitative results of VideoGLaMM on referring object segmentation**.

D. Ethics and Societal Impact

Our proposed dataset utilizes video samples from existing datasets which are released under the open-source public license and do not pose any privacy concerns. To the best of our knowledge, the dataset does not portray any strong biases or discrimination. We urge for the responsible use of our dataset and model, promoting research progress while safeguarding privacy.