HyperSeg: Hybrid Segmentation Assistant with Fine-grained Visual Perceiver

Supplementary Material

A. Additional Implementation Details

A.1. Evaluation Metrics

In our experiments, we use the widely used metrics to evaluate the performance of our HyperSeg on various segmentation tasks consistent with previous studies. Specifically, cumulative Intersection-over-Union (cIoU) for referring expression segmentation (RES), interactive segmentation, and generalized referring expression segmentation (G-RES), cIoU and the average of all per-image Intersection-over-Unions (gIoU) for reasoning segmentation task, region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ for reasoning video object segmentation (ReasonVOS), video object segmentation (VOS), referring video object segmentation (R-VOS), panoptic quality (PQ), mean intersection-over-Union (mIoU) for image generic segmentation, and mean average precision (mAP) for video instance segmentation (VIS).

A.2. Training Details

In our experiments, we use Phi-2 [20] with 2.7B parameters as our Large Language Model, SigLIP [56] as our vanilla encoder, and Swin-B [32] as our pyramid encoder. We use PyTorch to implement our HyperSeg and use Deepspeed zero-1 optimization for efficient training. Furthermore, the vanilla encoder and pyramid encoder are kept frozen, the LLM is finetuned with LORA (rank=8), the FVP, HER, and segmentation predictor are fully trained. Our codes and model weights will be publicly released.

A.3. Details about the Temporal Adapter

As shown in the Fig. 5, we inject the local information of the previous frame into the current frame through the Local Injection process and aggregate global prompt information of all the past t frames through the Global Aggregation.

B. Additional Experimental Results

B.1. Multi-modal Question Answering Benchmarks

Our HyperSeg is the first VLLM-based universal segmentation model for pixel-level image and video perception with complex reasoning and conversation capabilities, which is capable of tackling vision-language comprehension tasks. Therefore, we evaluate our model on various Multi-modal question answering benchmarks. As shown in Tab. 10, our HyperSeg achieves comparable performance compared with previous VLLMs like InstructBLIP [10], Qwen-VL [2], and LLaVA-1.5 [29] with fewer model parameters, demonstrating the insights into the model's powerful conversational and reasoning capabilities.

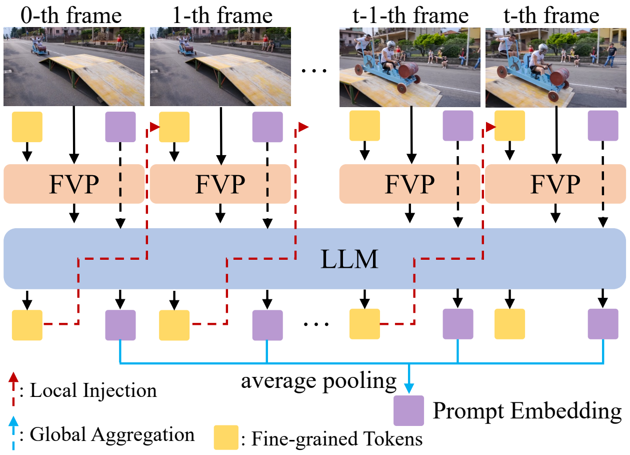


Figure 5. More Details about the Temporal Adapter.

Table 10. Quantitative results of our HyperSeg on Multi-modal question answering benchmarks. HyperSeg achieves promising performance compared with previous VLLMs in several widely used Multi-modal benchmarks.

Method	LLM	MMB	VQA^{v2}	GQA	POPE	SQA
BLIP-2 [23]	Vicuna-13B	-	65.0	41.0	85.3	61.0
InstructBLIP [10]	Vicuna-7B	36.0	-	49.2	-	60.5
InstructBLIP [10]	Vicuna-13B	-	-	49.5	78.9	63.1
Shikra [5]	Vicuna-13B	58.8	77.4	-	-	-
Qwen-VL [2]	Qwen-7B	38.2	78.8	59.3	-	67.1
Qwen-VL-Chat [2]	Qwen-7B	60.6	78.2	57.5	-	68.2
LLaVA-1.5 [29]	Vicuna-7B	64.3	78.5	62.0	85.9	66.8
HyperSeg	Phi-2-2.7B	67.9	78.2	60.9	86.6	66.2

B.2. Interactive Segmentation

We also evaluate HyperSeg on the COCO-Interactive validation set for the interactive segmentation task. As shown in Tab. 11, our HyperSeg achieves promising performance on various visual prompt types. Notably, our model surpasses previous segmentation specialists such as SAM [21], which utilizes a larger vision backbone and much more high-quality training data, and SEEM [64]. However, the VLLM-based model PSALM [59] exhibits superior performance in the interactive segmentation task. We hypothesize that this discrepancy arises from differences in feature scale utilization during the visual prompt sampling process: PSALM [59] employs the visual prompt features derived from a high-resolution Swin-based vision encoder, whereas HyperSeg utilizes features from a more streamlined CLIP-based visual encoder.

Table 11. Quantitative results on COCO-Interactive benchmark.

Method	Backbone	Box	Scribble	Mask	Point
SAM [21]	ViT-B	68.7	-	-	33.6
SAM [21]	ViT-L	71.6	-	-	37.7
SEEM [64]	DaViT-B	42.1	44.0	65.0	57.8
PSALM [22]	Swin-B	80.9	80.0	82.4	74.0
HyperSeg	Swin-B	77.3	75.2	79.5	63.4

C. Comparison of different settings

We also make setting comparisons between different models and our HyperSeg. As shown in Tab. 12, HyperSeg can handle more comprehensive segmentation tasks than previous segmentation specialists and MLLM-based methods. Firstly, HyperSeg can tackle both image-level and video-level perception tasks in one model enjoying the benefits of multi-task joint training. Secondly, HyperSeg performs various segmentation tasks, including long-text prompted referring and reasoning segmentation, category prompted generic segmentation, visual prompted interactive segmentation, and open-vocabulary segmentation.

D. Qualitative Results

In this section, we present more qualitative results to better demonstrate the segmentation capabilities of our HyperSeg involving various tasks in image and video domains.

D.1. Referring Expression Segmentation (RES)

Fig. 6 shows the visualization of HyperSeg on referring segmentation benchmarks (RefCOCO/+/g). Our model can effectively grasp the true meaning conveyed by the referring text and provide accurate pixel-level segmentation masks.

D.2. Interactive Segmentation

Fig. 7 presents the effectiveness of our HyperSeg in understanding the visual prompt and outputting the corresponding segmentation masks for the interactive segmentation tasks.

D.3. Panoptic Segmentation

Fig. 8 shows the qualitative results of HyperSeg in panoptic segmentation tasks, which needs both semantic and instance level dense predictions.

D.4. Reasoning Segmentation

Fig. 9 presents the effectiveness of our HyperSeg in understanding the complex question and perform segmentation according to the reasoning process.

D.5. Reasoning Video Object Segmentation (ReasonVOS)

Fig. 10 shows the effectiveness of HyperSeg in comprehending both the reasoning questions and temporal coherence. HyperSeg is capable of producing segmentation masks that maintain consistency across temporal sequences.

D.6. Video Object Segmentation (VOS)

The qualitative results of our method, HyperSeg, are illustrated in Fig. 11, demonstrating its capability in interpreting the visual prompt, provided by the ground truth object masks of the first frame, and producing accurate segmentation masks that maintain temporal consistency.

D.7. Video Instance Segmentation (VIS)

Fig. 12 illustrates the effectiveness of HyperSeg in performing instance-level video segmentation with class prompts, and executing accurate segmentation with instance tracking throughout the entire video.

Table 12. The comparison of different settings between our model and previous segmentation specialists and VLLM-based segmentation methods. Generic Seg denotes common class-based segmentation, such as panoptic segmentation and semantic segmentation. Open-set denotes the open-vocabulary segmentation. HyperSeg can perform more comprehensive segmentation tasks in one model.

Type	Method	Multi-task Training	Visual Type		Task Type				
			Image-level	Video-level	Referring Seg	Reasoning Seg	Generic Seg	Interactive Seg	Open-set
Segmentation Specialist	Mask2former [7]		✓				✓		
	OneFormer [19]		✓				✓		
	VLT [11]		✓		✓				
	LAVT [53]		✓		✓				
	PolyFormer [30]		✓		✓				
	ReferFormer [47]			✓	✓				
	OnlineRefer [46]			✓	✓				
	SEEM [64]	✓	✓	✓	✓		✓	✓	✓
	UNINEXT [25]	✓	✓	✓	✓		✓	✓	✓
	OMG-Seg [24]	✓	✓	✓			✓	✓	✓
VLLM-based Segmentation Network	LISA [22]	✓	✓		✓	✓			
	PixelLM [41]	✓	✓		✓	✓			
	GSVA [49]	✓	✓		✓				
	LaSagnA [45]	✓	✓		✓	✓	✓		
	OMG-LLaVA [58]	✓	✓		✓		✓		
	PSALM [59]	✓	✓	✓	✓		✓	✓	✓
	VISA [51]	✓	✓	✓	✓	✓			
	HyperSeg (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

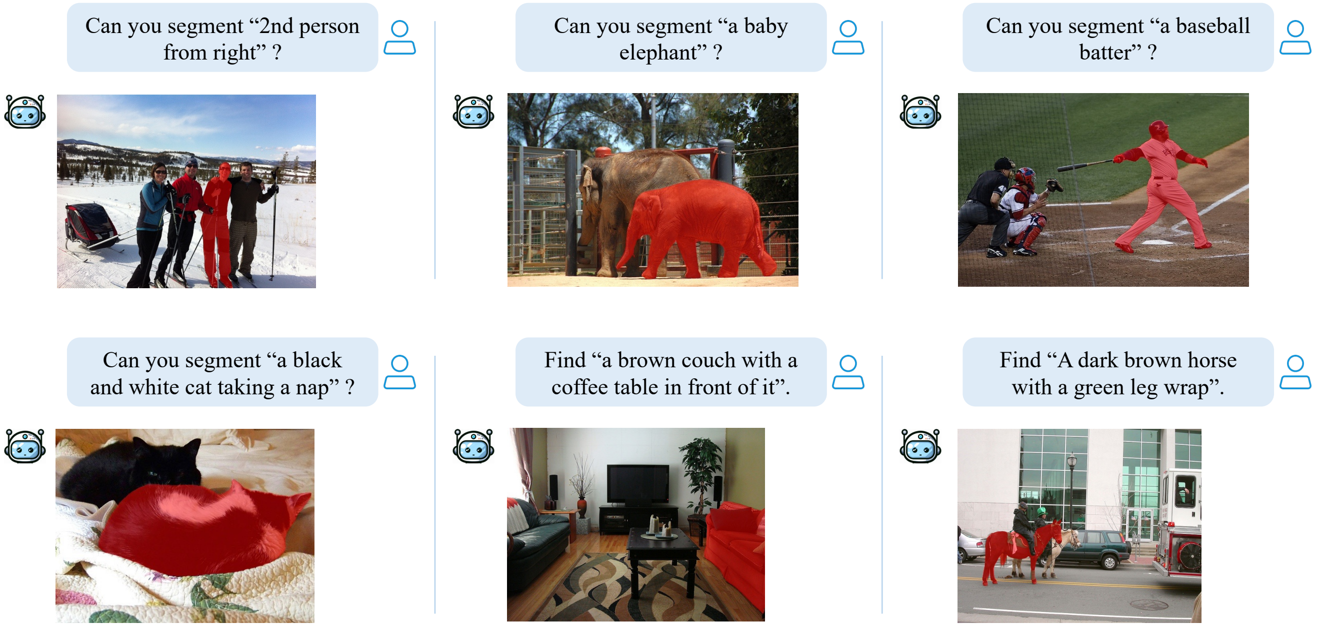

Can you segment "2nd person from right" ?
Can you segment "a baby elephant" ?
Can you segment "a baseball batter" ?
Can you segment "a black and white cat taking a nap" ?
Find "a brown couch with a coffee table in front of it".
Find "A dark brown horse with a green leg wrap".

Figure 6. Qualitative results of HyperSeg's capability in referring expression segmentation.

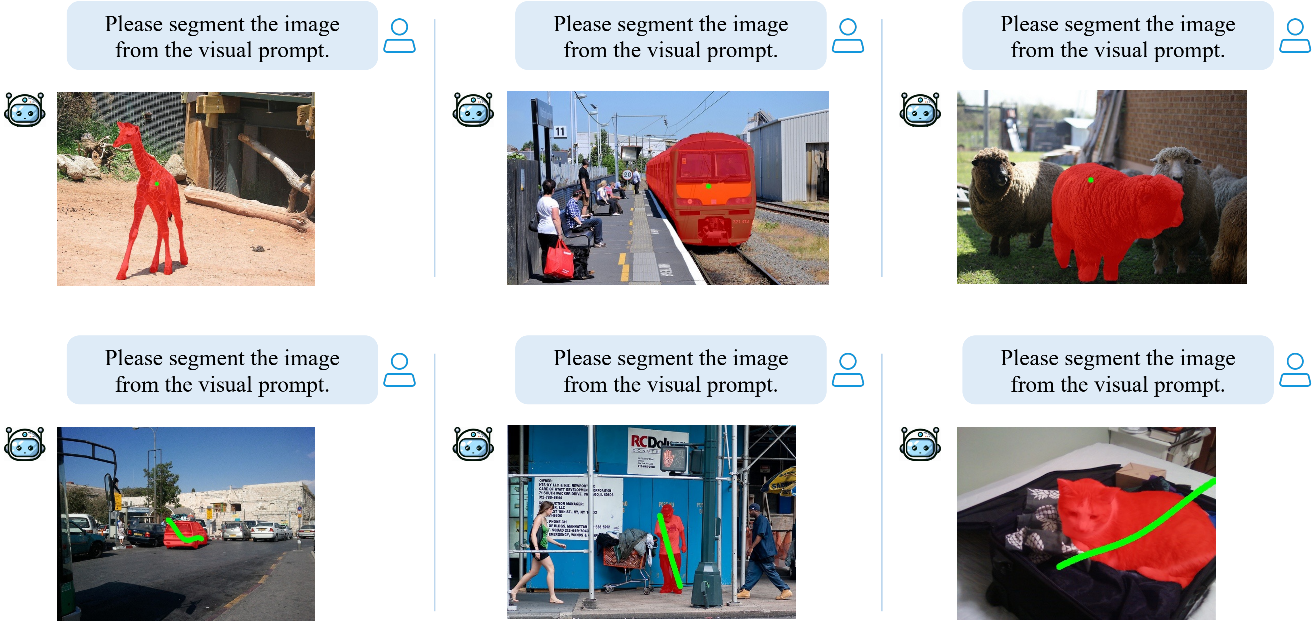

Please segment the image from the visual prompt.
Please segment the image from the visual prompt.
Please segment the image from the visual prompt.
Please segment the image from the visual prompt.
Please segment the image from the visual prompt.
Please segment the image from the visual prompt.

Figure 7. Qualitative results of HyperSeg in interactive segmentation. The green marker indicates the provided visual prompts, such as point and scribble.

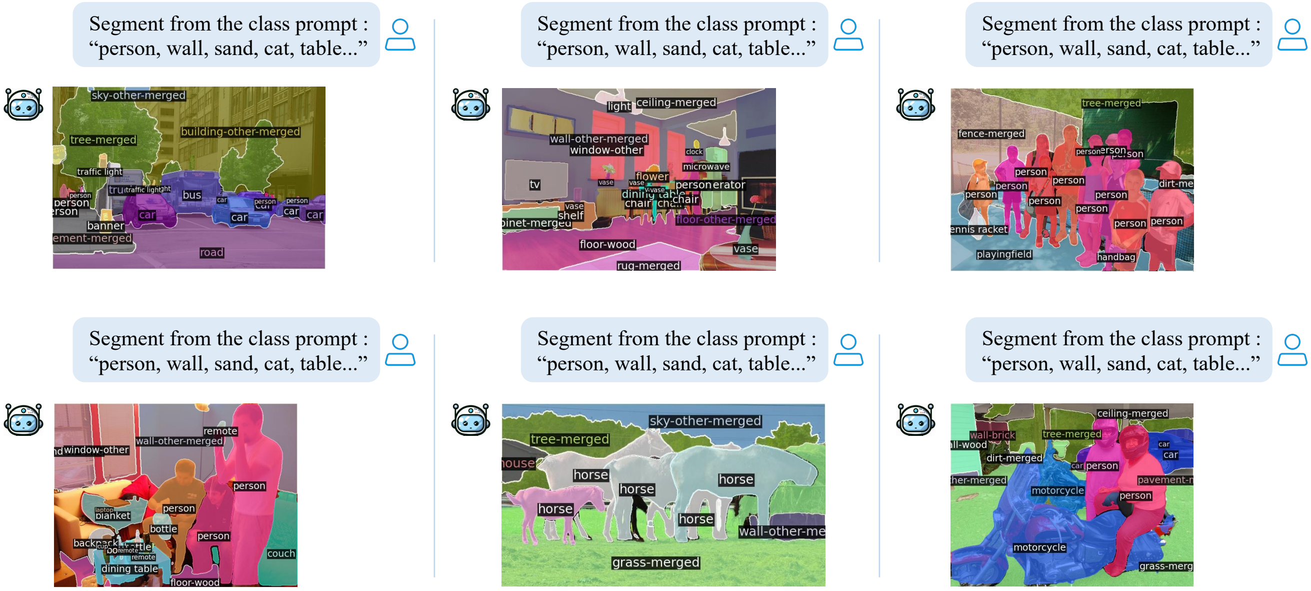
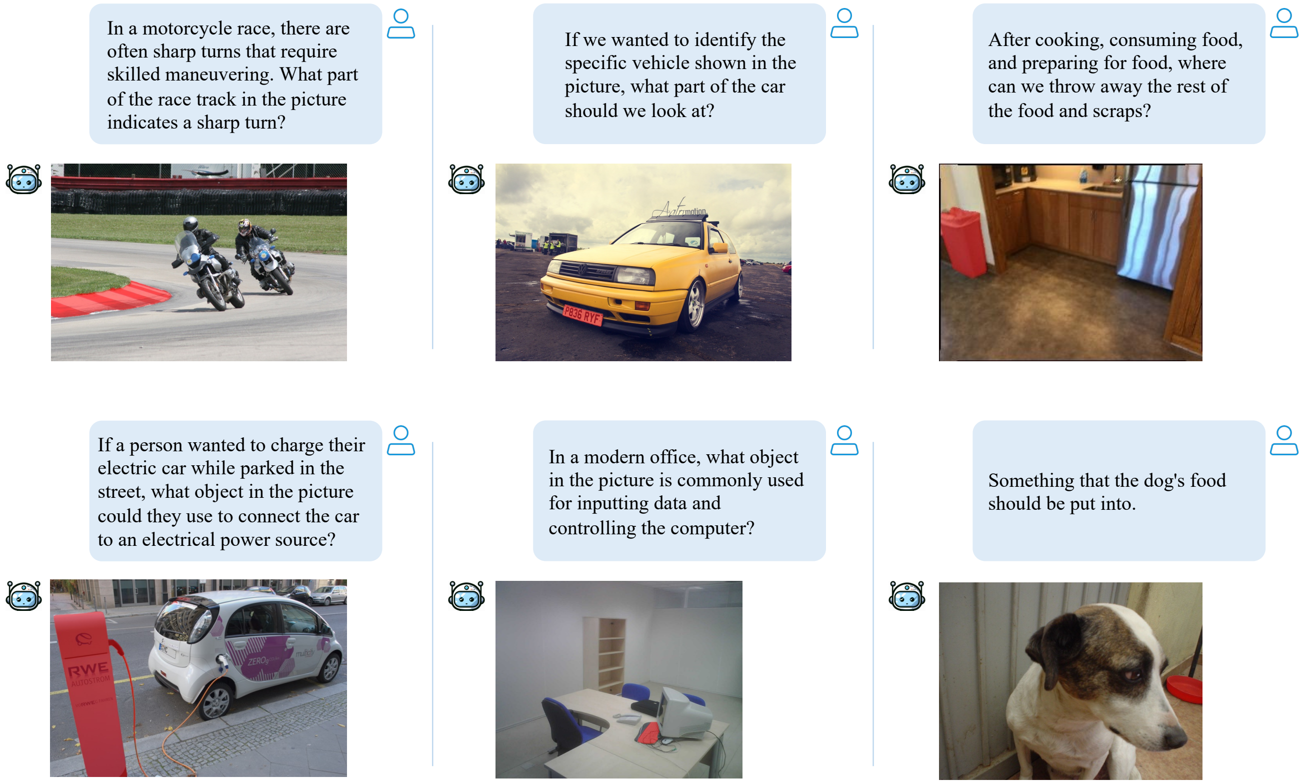


Figure 8. Qualitative results of HyperSeg in panoptic segmentation.



In a motorcycle race, there are often sharp turns that require skilled maneuvering. What part of the race track in the picture indicates a sharp turn?

If we wanted to identify the specific vehicle shown in the picture, what part of the car should we look at?

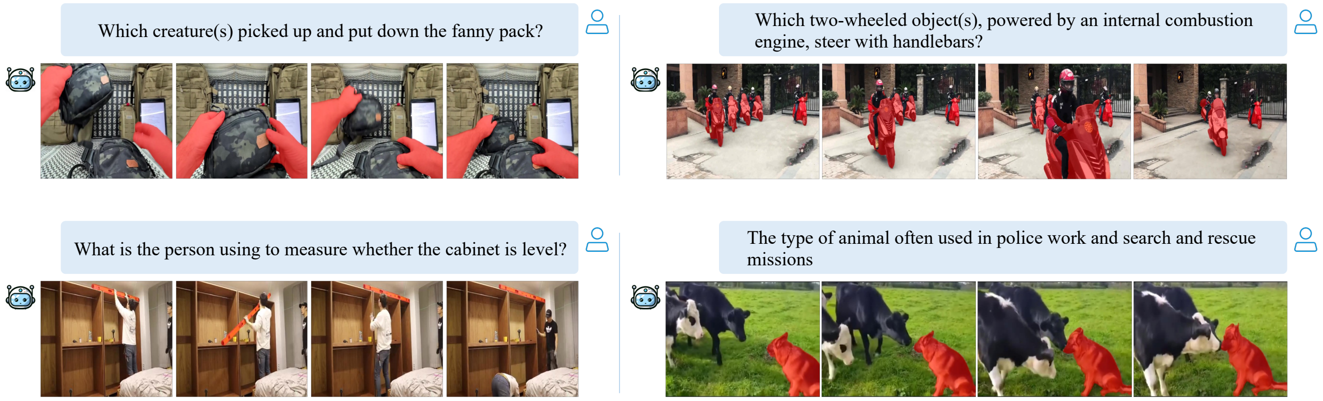
After cooking, consuming food, and preparing for food, where can we throw away the rest of the food and scraps?

If a person wanted to charge their electric car while parked in the street, what object in the picture could they use to connect the car to an electrical power source?

In a modern office, what object in the picture is commonly used for inputting data and controlling the computer?

Something that the dog's food should be put into.

Figure 9. Qualitative results of HyperSeg in reasoning segmentation.



Which creature(s) picked up and put down the fanny pack?

Which two-wheeled object(s), powered by an internal combustion engine, steer with handlebars?

What is the person using to measure whether the cabinet is level?

The type of animal often used in police work and search and rescue missions

Figure 10. Qualitative results of HyperSeg demonstrate its capability in the complex reasoning video object segmentation task, effectively managing challenging video data and producing temporally consistent results following the reasoning process.

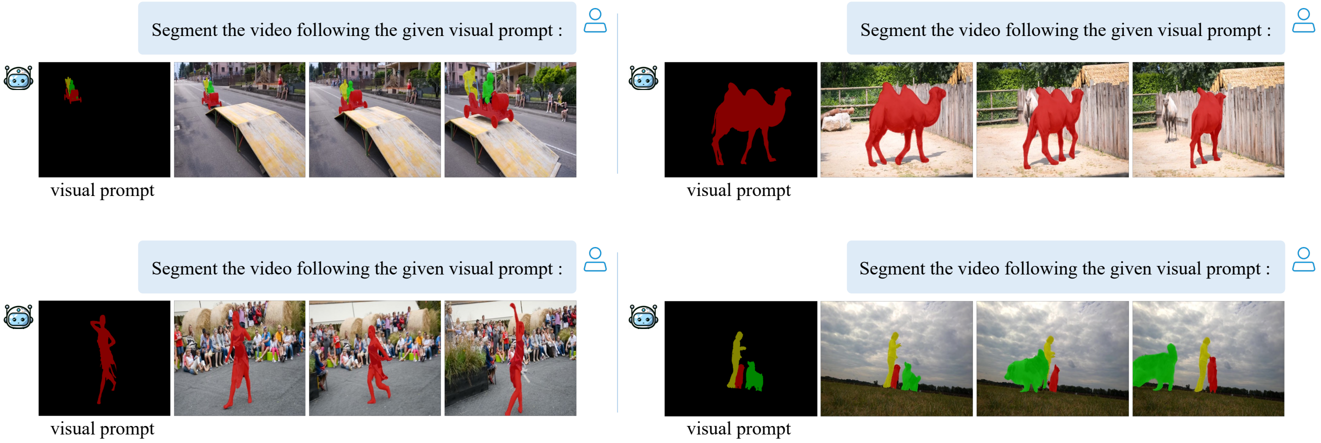

Segment the video following the given visual prompt :
visual prompt
Segment the video following the given visual prompt :
visual prompt
Segment the video following the given visual prompt :
visual prompt
Segment the video following the given visual prompt :
visual prompt

Figure 11. Qualitative results of HyperSeg in semi-supervised video object segmentation tasks. With the visual prompts provided by the ground truth object masks of the first frame, HyperSeg demonstrates its ability to achieve accurate segmentation while maintaining temporal consistency.

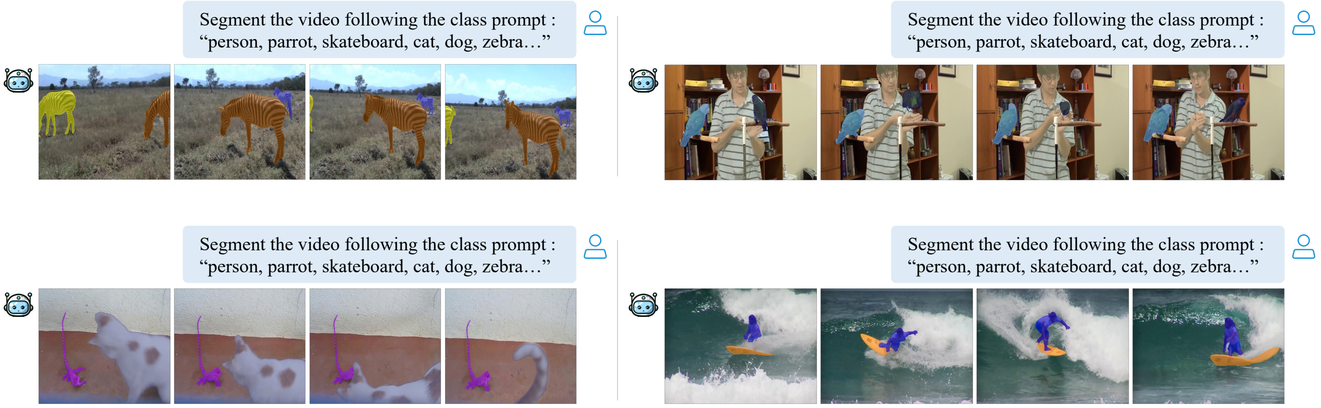

Segment the video following the class prompt : “person, parrot, skateboard, cat, dog, zebra…”
Segment the video following the class prompt : “person, parrot, skateboard, cat, dog, zebra…”
Segment the video following the class prompt : “person, parrot, skateboard, cat, dog, zebra…”
Segment the video following the class prompt : “person, parrot, skateboard, cat, dog, zebra…”

Figure 12. Qualitative results of HyperSeg in video instance segmentation tasks. Utilizing the class text prompts and instance tracking strategies, HyperSeg exhibits its capability to achieve precise segmentation while ensuring temporal consistency.