

UFVideo: Towards Unified Fine-Grained Video Cooperative Understanding with Large Language Models

Supplementary Material

A. Training Recipe

The complete list of training datasets for UFVideo is detailed in Tab. 8. Utilizing a two-stage training approach, we initially pre-train the model to integrate the tasks of temporal grounding and video segmentation, addressing the limitations of the VideoRefer-7B model [84] which lacks these capabilities. We draw upon the temporal training datasets from LLaVA-ST [26] and Distime [86], adapting them with relative temporal tokens to better suit our model’s requirements. For video segmentation, we use commonly employed video segmentation training datasets to synchronize the LLM output embedding with the mask decoder. In the second stage, we compile a hybrid large-scale dataset encompassing global video question answering, pixel-level video object referring, video segmentation, temporal video grounding, and self-constructed multi-grained video cooperative understanding tasks. This dataset contains over 3 million samples, and we plan to open source it soon!

B. UFVideo-Dataset

B.1. Construction Details

The training dataset of PixRQA and PixHQA are constructed using object-level QA data from the VideoRefer-700K dataset [84]. The benchmark of PixRQA is built from the VideoRefer-Bench-D dataset and PixHQA is built from the VideoRefer-Bench-Q dataset. The VideoRefer-700K dataset includes masks for referring objects in corresponding frames. For our task definition, in PixRQA, we randomly select a mask from one of the initial frames as the visual prompt, with the remaining frames serving as label masks. In PixHQA, all frame masks are used as label masks. Since the frames with masks in the VideoRefer data are not continuous, but PixTRQA requires answering the corresponding time and masks within that range, continuous frames must be maintained. Therefore, the training dataset and benchmark of PixTRQA are based on the ReVOS [78] training set and validation set, respectively., which involves segmentation based on an implicit question, ensuring that the object only appears during the relevant time in the video. We also filter the ReVOS data, retaining samples with continuous frames and removing those with dispersed occurrences to ensure a complete time segment as the label. Thus, all three tasks retain corresponding masks and times, necessitating the reconstruction of semantic questions and responses. We selected the latest Qwen3VL-235B-A22B-Instruct model [61] for semantic question-answer construction.

B.2. UFVideo Dataset Prompt Template

We design three prompt templates to construct the training datasets for UFVideo’s three tasks. The detailed prompts are presented in Fig. 4.

B.3. Qwen3-VL Annotation Prompt Template

To annotate high-quality data, we selected the latest Qwen3-VL-235B-A22B-Instruct as our semantic annotation model, which the prompts shown in Fig. 5. For the input video and its corresponding object boxes, we generate short descriptions as prompts and long descriptions as semantic responses.

C. Experiments

C.1. Evaluation Metrics

For general video understanding tasks, as presented in Tab. 1, the metrics are present in order: Action Sequence (AS), Action Prediction (AP), Action Antonym (AA), Fine-grained Action (FA), Unexpected Action (UA), Object Existence (OE), Object Interaction (OI), Object Shuffle (OS), Moving Direction (MD), Action Localization (AL), Scene Transition (ST), Action Count (AC), Moving Count (MC), Moving Attribute (MA), State Change (SC), Fine-grained Pose (FP), Character Order (CO), Egocentric Navigation (EN), Episodic Reasoning (ER), Counterfactual Inference (CI).

For video object referring tasks, we evaluate on VideoRefer-Bench-D and VideoRefer-Bench-Q, where VideoRefer-Bench-D metrics are Subject Correspondence (SC), Appearance Description (AD), Temporal Description (TD) and Hallucination Detection (HD) and use GPT-4o-2024-08-06 to evaluate semantic scores from 0 to 5. VideoRefer-Bench-Q is multi-choice question, the metrics are choice accuracy, including four dimensions of understanding, including Basic Questions (BQ), Sequential Questions (SQ), Relationship Questions (RQ), Reasoning Questions (RQ) and Future Predictions (FP).

For video segmentation tasks, we use region similarity \mathcal{J} , contour accuracy \mathcal{F} , and their combination $\mathcal{J}\&\mathcal{F}$ as evaluation metrics.

For temporal video grounding tasks, the primary metrics are tIoU, R@0.3, R@0.5, and R@0.7, where tIoU measures the IoU between the predicted and ground truth time intervals, and R@k represents the proportion of tIoU values exceeding k.

For UFVideo-Bench, we adopt \mathcal{J} , \mathcal{F} , and $\mathcal{J}\&\mathcal{F}$ to measure segmentation performance, and SAvg. represents semantic average scores following VideoRefer-Bench-D evalu-

Table 8. The training recipe of UFVideo, which joint common video understanding tasks such as general video understanding, pixel-level video object referring, pixel-level video segmentation and temporal video grounding tasks, and multi-grained video cooperative understanding tasks.

Stage	Dataset	Task	#Samples	#Ratio
1	MeViS [111]	Segmentation	23K	0.5
	Ref-DAVIS17 [49]		0.6K	0.5
	Ref-YouTube-VOS [56]		13K	0.5
	ReVOS [78]		29K	0.5
	Anet-Caption [23]	Temporal	37K	0.5
	Anet-Caption-expand [86]		37K	0.5
	Anet-RTL [20]		10K	0.5
	COIN [59]		10K	0.5
	DiDeMo [1]		33K	0.5
	ET-Instruct [38]		72K	0.5
	Grounded-VideoLLM [62]		17K	0.5
	HiREST [85]		0.8K	0.5
	InterVid-G [66]		134K	0.5
	InternVid-TG [86]		86K	0.5
	Moment-10M [50]		60K	0.5
	QuerYD [44]		34K	0.5
	VTG-IT [16]		32K	0.5
	ViTT [21]		5K	0.5
	VTimeLLM [19]		139K	0.5
	YouCook2 [92]		1.8K	0.5
2	AGQA [15]	General QA	1600K	1
	CLEVRER [80]		82K	1
	EgoQA [14]		21K	1
	Intent-QA [28]		12K	1
	Kinetics [29]		26K	1
	LLaVA-Video-178K [90]		300K	1
	NEXT-QA [75]		34K	1
	ShareGPT4Video [8]		39K	1
	STAR [70]		45K	1
	SthSthV2 [42]		40K	1
	TextVR [73]		39K	1
	VCG-Plus-112K [41]		112K	1
	Videochat2-Conv [30]		9K	1
	Videochatgpt-100K [40]		100K	1
	WebVid [4]		400K	1
	WebVid-QA [79]		86K	1
	YouCook2 [92]	8K	1	
	MeViS [111]	Segmentation	23K	5
	Ref-DAVIS17 [49]		0.6K	10
	Ref-YouTube-VOS [56]		13K	7
ReVOS [78]	29K		5	
VideoRefer-700K (Detail) [15]	Referring	125K	2	
VideoRefer-700K (Short) [15]		50K	1	
VideoRefer-700K (QA) [15]		75K	2	
Anet-Caption [23]	Temporal	37K	1	
Anet-Caption-expand [86]		37K	1	
Anet-RTL [20]		10K	1	
COIN [59]		10K	1	
DiDeMo [1]		33K	1	
ET-Instruct [38]		72K	1	
Grounded-VideoLLM [62]		17K	1	
HiREST [85]		0.8K	1	
InterVid-G [66]		134K	1	
InternVid-TG [86]		86K	1	
Moment-10M [50]		60K	1	
QuerYD [44]		34K	1	
VTG-IT [16]		32K	1	
ViTT [21]		5K	1	
VTimeLLM [19]		139K	1	
YouCook2 [92]		1.8K	1	

ation settings. tIoU follows the same definition as in temporal grounding tasks. For PixHQA, we distinguish between temporal point-based and temporal interval-based answers, excluding temporal description metrics for point-based answers following VideoRefer-Bench-D evaluation. For Pix-

TRQA, we only compute segmentation mask metrics for frames within time ranges satisfying the tIoU threshold.

C.2. Qualitative Results

We present additional outputs from UFVideo for common video understanding tasks and multi-grained video cooperative understanding task. Multi-grained video cooperative understanding results of PixHQA, PixTRQA and PixRQA are presented in Fig. 6, Fig. 7 and Fig. 8, respectively. General video understanding results are shown in Fig. 9, video object referring results in Fig. 10, video segmentation results in Fig. 11 and Fig. 12, and temporal video grounding results in Fig. 13. These visualizations demonstrate that UFVideo achieves strong performance and flexibly handles multi-grained video understanding tasks.

C.3. More Experimental Results

As shown in Tab. 9, we present the complete experimental results on the ReVOS dataset for the reasoning video object segmentation task. The results demonstrate that UFVideo outperforms state-of-the-art baselines in both referring and reasoning segmentation, proving that UFVideo gains effective enhancement from multi-grained information.

D. Limitation & Future Work

In this work, we present UFVideo, which focuses on video-level multi-grained understanding but gives less attention to image-level understanding, which could provide additional enhancement for video comprehension through mutual reinforcement. In future work, we aim to develop a more comprehensive and powerful visual LLM to achieve superior performance on visual cooperative understanding tasks, better serving real-world applications.

Table 9. Experimental results on ReVOS dataset for reasoning video object segmentation task, and comparison with state-of-the-art methods. The red/blue indicates the best/second-best results.

Model	Size	Referring			Reasoning			Overall		
		\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$
<i>Specialist Models</i>										
MTTR [5]	–	29.8	30.2	30.0	20.4	21.5	21.0	25.1	25.9	25.5
LMPM [11]	–	29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4
ReferFormer [72]	–	31.2	34.3	32.7	21.3	25.6	23.4	26.2	29.9	28.1
<i>Multi-modal LLMs</i>										
LISA [24]	13B	45.2	47.9	46.6	34.3	39.1	36.7	39.8	43.5	41.6
TrackGPT [57]	13B	48.3	50.6	49.5	38.1	42.9	40.5	43.2	46.8	45.0
VISA [78]	13B	55.6	59.1	57.4	42.0	46.7	44.3	48.8	52.9	50.9
HyperSeg [68]	3B	56.0	60.9	58.5	50.2	55.8	53.0	53.1	58.4	55.7
InstructSeg [69]	3B	54.8	59.2	57.0	49.2	54.7	51.9	52.0	56.9	54.5
GLUS [35]	7B	56.0	60.7	58.3	48.8	53.9	51.4	52.4	57.3	54.9
ViLLa [91]	6B	–	–	–	–	–	–	54.9	59.1	57.0
Sa2VA [82]	4B	–	–	–	–	–	–	–	–	53.2
RGA3 [63]	7B	58.7	62.3	60.5	53.1	57.7	55.4	55.9	60.0	58.0
UniPixel [39]	7B	<u>64.2</u>	<u>68.5</u>	<u>66.4</u>	<u>59.6</u>	63.9	61.8	<u>61.9</u>	<u>66.1</u>	<u>64.0</u>
UFVVideo	7B	65.4	69.8	67.6	59.8	<u>63.8</u>	61.8	62.7	66.9	64.8

Box 1: Prompt Template for PixRQA, PixHQA and PixTRQA tasks.

PixRQA: (joint general question answering, video object referring and video segmentation)

Question: If object₁ <region> short description, object₂ <region> short description, ..., object_n <region> short description, what is a likely future event? And please generate the mask in every frames.

Answer: Long description. The segmentation mask: object₁[SEG], object₂[SEG], ..., object_n[SEG].

PixHQA: (joint general question answering, referring video segmentation and temporal grounding)

Question for timepoint: What object₁ short description, object₂ short description, ..., object_n short description are doing in the <Temp-x>, and generate the masks?

Question for time period: What object₁ short description, object₂ short description, ..., object_n short description are doing in the {<Temp-x1><Temp-x2>}, and generate the masks?

Answer: Long description. The segmentation mask: object₁[SEG], object₂[SEG], ..., object_n[SEG].

PixTRQA: (joint general question answering, referring video segmentation and temporal retrieval)

Question: What object₁ <region> short description, object₂ <region> short description, ..., object_n <region> short description are doing? And please generate the time period and object mask.

Answer: The Time is {<Temp-x1><Temp-x2>}. Long description. The segmentation mask: object₁[SEG], object₂[SEG], ..., object_n[SEG].

Figure 4. Prompt template for three UFVVideo-Bench tasks. <Region> denotes as the special token for video object referring task. [SEG] denotes as the special token for video segmentation, <Temp-x> represents the special token for temporal video grounding.

Box 2: Prompt template for annotation in Qwen3-VL-235B-A22B-Instruct

Question :

I upload multiple frames of the video as single-frame images, arranged in the original video order. Next, I upload the box positions of single or multiple objects in each frame as lists. Please provide a <Short Description> of about 10 words for each object, and a <Long Description> of all objects and their relationships. Both descriptions apply to the entire video, not individual frame. In the <Long Description>, use the object nouns in the <Short Description> to refer to each object. All descriptions should be in English.

Object box lists: box_list_prompt.

Provide your description strictly in the following format: <Short Description>:{object short description prompt} <Long Description>: your long description of all objects and their relationships.

Answer Examples:

Single object answer:

<Short Description>: object_1: Blue monster truck mid-air during stunt jump

<Long Description>: The blue monster truck, object_1, is captured mid-air, performing a high jump over a dirt track. In the background, several other monster trucks are visible on the ground, including a white one and additional blue ones, positioned along the track as if awaiting their turn or observing the stunt. The scene takes place in an outdoor arena with spectators behind barriers and trees and buildings in the distance, suggesting a live event or competition. The blue monster truck's shadow is cast on the ground below, emphasizing its height and motion.

Multi objects answer:

<Short Description>: object_1: Blue monster truck soaring through the air. object_2: Line of monster trucks on dirt track.

<Long Description>: The blue monster truck, object_1, is captured mid-air, performing a high jump over the dirt track. Below it, object_2, a line of monster trucks, is parked or moving slowly along the track, creating a dynamic contrast between the airborne vehicle and the grounded ones. The scene unfolds at an outdoor event with spectators visible in the background, emphasizing the spectacle of the stunt.

Figure 5. Prompt template for annotation in Qwen3-VL-235B-A22B-Instruct. We present the question template and answer examples with single or multi objects.



Q: What is **object 1**: silver SUV driving on mountain road curve doing in the 3.2s ?

A: A silver SUV cruises along a winding mountain road, flanked by green meadows and evergreen forests. Towering rocky peaks rise behind it under a partly cloudy sky. Moving at a relaxed pace, the vehicle follows the curve with its front facing the viewer, bordered by a wire fence on one side. It's a peaceful, scenic drive—harmonizing machine and nature in a remote alpine setting.



Q: What are **object 1**: pedestrian walking close to the wall and **object 2**: man in gray shirt crossing zebra crossing doing in the 8.6s ?

A: The pedestrian just appeared in the video, walking slightly ahead and to the right of the man in the white shirt. They're not interacting—just two strangers crossing together by coincidence, moving in parallel with no connection beyond sharing the same crosswalk for a moment. One in dark clothes, one in light—they're anonymous figures passing through the city, side by side but worlds apart.



Q: What is **object 1**: large brown bear doing in the 1.5s to 6.9s ?

A: The large brown bear is walking steadily across a rocky, sunlit enclosure. Its thick, shaggy fur glistens slightly under the sunlight as it moves forward with a deliberate gait. The bear navigates around scattered boulders and steps over uneven ground, with its head lowered as if sniffing or inspecting the terrain. Behind it, a textured stone wall and patches of green foliage frame the scene, suggesting a zoo or wildlife sanctuary habitat. The bear remains the sole focus throughout, with no other animals or human elements visible, emphasizing its solitary presence and naturalistic behavior within the constructed environment.



Q: What are **object 1**: person in wheelchair with colorful striped socks and **object 2**: gray dog with white markings doing in the 4.2s to 8.1s ?

A: A gray dog with white markings walks on the paved ground, moving toward the camera and occasionally licking its lips. Nearby, a person in a wheelchair, wearing a teal shirt and colorful striped socks, remains seated and stationary throughout the sequence. The wheelchair is positioned on a tiled walkway adjacent to a grassy area with a tree, and the dog approaches from the grassy side, creating a dynamic interaction between the stationary person and the moving dog.

Figure 6. Visualized results of PixHQA.



Q: What is *object_1*: man throwing frisbee on sandy beach doing and when in the video ?

A: The time is 0s to 3.4s in the video. A man is standing on a sandy beach, preparing to throw a frisbee. He is shirtless, wearing dark shorts, and appears to be in motion as he extends his arm to launch the frisbee into the air. The frisbee is visible in mid-flight in some frames, indicating the action sequence. In the background, there is a calm sea with sailboats and a distant shoreline with structures. The sky is overcast, and part of a beach umbrella is visible on the right edge of the frame, suggesting a leisurely beach setting. The man's posture and the trajectory of the frisbee convey a casual, recreational moment.



Q: What is *object_1*: person walking away holding a red fire extinguisher doing and when in the video ?

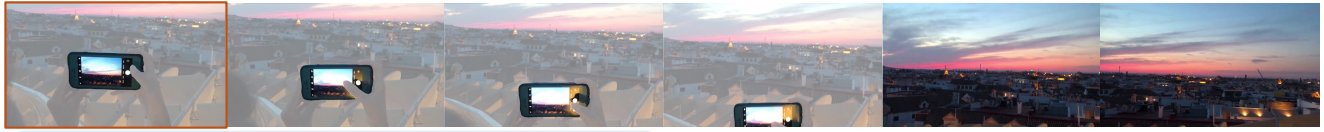
A: The time is 3.7s to 6.2s in the video. Two large model fighter jets are positioned on a grassy airfield, appearing to have just landed or crashed. Two men, one in a black jacket with a green hood and another in a dark jacket with a red scarf, are inspecting and handling the models, adjusting their positions and checking for damage. A third person, *object_1*, is seen walking away from the scene while holding a red fire extinguisher, suggesting a safety precaution or response to a minor incident. The background shows an airport environment with hangars and distant buildings under an overcast sky. The interaction between the men and the models indicates they are likely hobbyists or operators of the remote-controlled aircraft, engaged in post-flight assessment or recovery.



Q: What is *object_1*: pink-sleeved person holding a mango near tree doing and when in the video ?

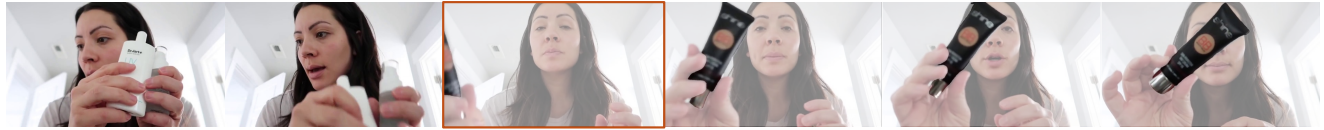
A: The time is 5.4s to 9.5s in the video. A person wearing a pink sleeve is holding a large, green-and-yellow mango, gently rotating it in their hand while standing next to a mango tree. The tree's thick trunk and leafy branches form the background, with a blue pole visible nearby. The person's face is partially visible, showing red lipstick and a smile, suggesting enjoyment or pride in the fruit. The mango appears ripe with a slight blemish, and the person's careful handling indicates inspection or preparation for harvest. The scene is outdoors in natural daylight, conveying a tranquil, garden-like atmosphere.

Figure 7. Visualized results of PixTRQA.



Q: Please describe what is the object_1 <region> doing in the video ?

A: This is an iphone that is being held in both hands to take a picture of a city sunset, the sunset is pink and very beautiful. The shooter is adjusting the focus.



Q: Please describe what is the object_1 <region> doing in the video ?

A: The item referred to was BB cream. It's a tube of cream with a black body and a silver cap, and the label says 'shine' and 'BB'. The lady picks up the BB cream and shows it to the camera.



Q: Please describe what is the object_1 <region> doing in the video ?

A: The screwdriver in the video is green with a black handle. It is being held in someone's hand.

Figure 8. Visualized results of PixRQA.



Q: Can you describe the scene of the video in detail ?

A: Yes! This video shows a football match between a team wearing blue and yellow and a team wearing red, with a referee present. The soccer field is green with white lines, and there is a blue and white advertisement banner in the background. In some frames, there are also people sitting on benches on the sides of the field.

Q: Are there any notable interactions between the players and the referee ?

A: Yes, there are several interactions between the players and the referee. The players often look towards the referee for guidance or to appeal for calls. The referee is also seen pointing towards different areas of the field at times, presumably to explain certain calls or decisions.

Q: Can you describe the actions of the man playing soccer ?

A: The man playing soccer on the field is constantly moving and kicking the ball around. He is also interacting with other players on the field, passing the ball to them or trying to steal it away. At times, the man is also being watched by the referee, who is keeping a close eye on the game.

Figure 9. Visualized results of multi-turn open-ended video QA of UFVideo.



Q: Please give a detailed description of the highlighted object <region> in the video.

A: The panda in the video has a predominantly white body with black patches around its eyes, ears and limbs. It has a round and fluffy appearance with contrasting black and white colors. The panda's fur looks soft and thick, giving it a plush feel. Its eyes are small and black and its ears are round and black. The panda's overall demeanor is calm and relaxed. The panda first authors sitting and then lying down.



Q: How does the movement of <object1><region> relate to the movement of <object2><region> ?

- (A) <object1> is stationary while <object2> is moving.
- (B) Both <object1> and <object2> are navigating a dirt track.
- (C) <object1> is moving in the opposite direction of <object2> .
- (D) <object1> and <object2> are not related.

A: (B) Both <object1> and <object2> are navigating a dirt track.

Figure 10. Visualized results of VideoRefer-Bench-D (top) and VideoRefer-Bench-Q (bottom).



Q: Please segment the elephant facing us and then turning around to walk away ?



Q: Please segment the plane facing our direction and moving ?



Q: Please segment the two moving model airplanes ?

Figure 11. Visualized results of MeViS.



Q: Please segment which bony fish(es) belongs/belong to the genus *Carassius* of the family *Cyprinidae* of the order *Cypriniformes*?



Q: Please segment which tiger(s) is/are shown with only its head visible and not its body in the scene?



Q: Please segment the person who stays close to the white wall throughout the scene, moving back and forth.



Q: Please segment which cat(s) have entirely black fur?

Figure 12. Visualized results of ReVOS. The first two rows illustrate the reasoning process of ReVOS, while the last two rows demonstrate the referring aspect of ReVOS.

