SF^{2}T: Self-supervised Fragment Finetuning of Video-LLMs for Fine-Grained Understanding

Supplementary Material

In this supplementary material, Section A presents SF^{2}T's performance on video caption tasks and additional exemplary visualizations of the attention map, while Section B provides more details about FineVidBench.

A. More Results and Cases

In addition to FineVidBench and public video understanding benchmarks, we also evaluated the video caption task (Table 1) using GPT-4o mini, assessing fluency, relevance, informativeness, and correctness, with a maximum score of 40. The results show that incorporating SF^{2}T improves performance, highlighting that fine-grained understanding also benefits video captioning. However, after fine-tuning, MiniCPM-V 2.6 produced shorter responses, leading to a decrease in its informativeness score.

Methods	LLaVA-NeXT -Video	MiniCPM-V 2.6	VideoLLaMA 2.1	Qwen2 -VL
Base	33.20	32.61	22.53	29.76
Base+SF^{2}T	**33.29**	29.73 ↓	**30.99**	**30.05**
Base(SFT)	27.62	29.60	27.19	29.66
Base(SFT)+SF^{2}T	**30.50**	**31.31**	**28.94**	**31.04**

Table 1. Performance on video caption task. The results show that incorporating SF^{2}T yields higher scores (except MiniCPM-V 2.6), likely due to its enhanced temporal sensitivity and understanding.

As shown in Figure 1, we present more attention maps for Qwen2-VL on the Action task, focusing on cases where the model's predictions were corrected after applying SF^{2}T.

B. Details of FinevidBench

B.1. Question-Answer Templates

Table 2 delineates the question templates for each task. For the answers, Scene-level tasks include Action task, which are composed of the "visual synonyms" and other verbs; Effect task, which are scripted by researchers based on video content; and Speed task, which offer fixed options: fast, slow, normal, and no speed. Fragment-level tasks encompass Frame Count, with answers ranging from 2 to 6; Meaning of Order, using ordinal numbers as responses; Frame Comparison and Adjust or Not, with responses of Yes, No, and Not sure; and Rearrangement, where the answer is a permutation of N numbers, with N representing the number of input frames. The Question-Answer database is generated through a process of template creation followed by iterative refinement using GPT-4. For Action and Effect tasks, each original video is queried three times using different question formulations. For Speed tasks, one query is conducted for both the original and the speed-altered versions of the video. For Fragment-Level tasks, all five questions are posed for each unique frame count.

B.2. Detailed Results

• Scene Level

Table 3 illustrates the types of action effects and examples in the Effect tasks. For the affected objects, common physical attributes and quantities of objects are considered; notably, the positional relationship, spatial distance, and similarity between two objects are examined. Regarding action attributes, the intensity and completeness of the action are evaluated. Special actions include slight movement, multiple-object movements where several affected objects undergo motion, and compound movements involving two or more atomic actions linked in time. Additionally, camera movements and the inclination of the surface on which objects move are assessed. Table 4 presents the results categorized under the Effect classification. Overall, models performed well in Physical Attributes and Action Intensity, likely due to the ability to infer such information by comparing images before and after the action occurs. However, models exhibited subpar performance in Action Completion and Camera Motion. The former suggests a lack of understanding regarding the distinction between completed and incomplete actions in terms of their effects, while the latter is attributable to the inherent variability and complexity of camera movements. For other tasks, the majority of models exhibited moderate performance.

• Fragment Level

Table 5 presents the results for all tasks in the fragment level under varying input frame counts. From the results, we can observe that except for Video-CCAM, the models' ability to count frames significantly declines as the frame count increases. Regarding the understanding of order concepts, most models show a clear upward trend, except for ShareGPT4Video. Models generally perform well on the frame comparison task, likely due to extensive training with image-text pairs. Since the input consistently involves two frames, the results show no significant variation, as expected. For Rearrangement, all results hover around random values, suggesting that while models recognize incorrect sequence orders, they cannot correct them, indicating a failure to grasp the dynamic processes of videos truly.

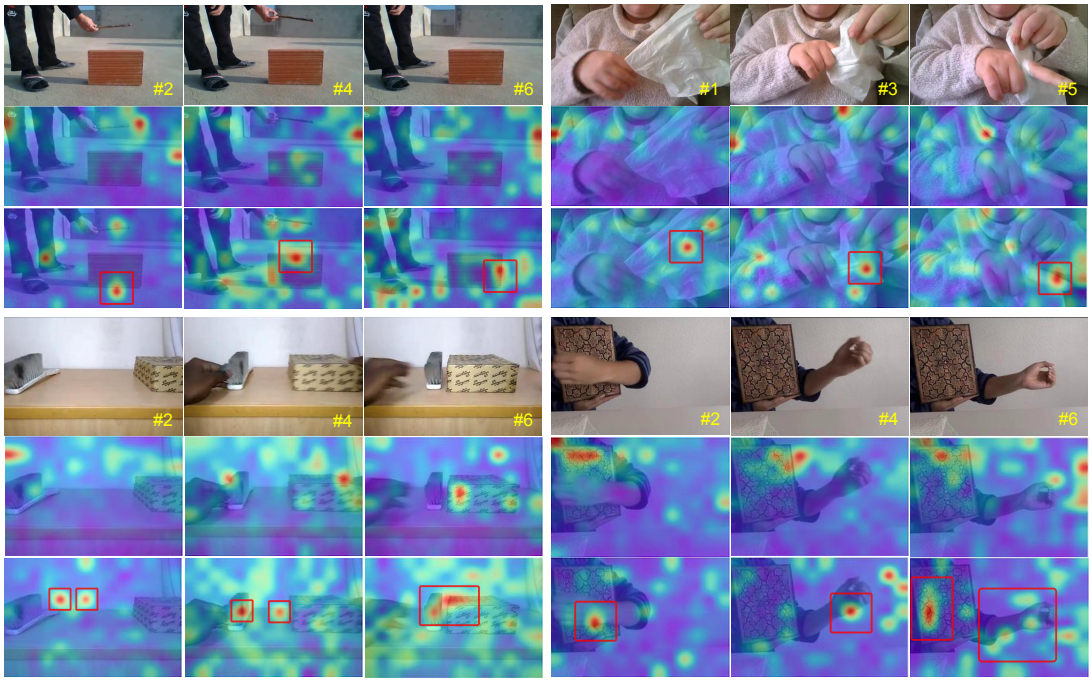

Figure 1. Four exemplary visualizations of the attention map on Qwen2-VL. For each example: top - Original frames; middle - Base (SFT); bottom - SF^{2}T applied. As highlighted by the red boxes, applying SF^{2}T enables the model to better focus on action execution areas and interacting objects, while also predicting the direction of motion.

Tasks		Question
Scene Level	Action	Which activity can be seen in the video?
	Effect	After the action takes place, what changes occur to the object?
		During the process of the action, what changes occur to the object?
		After the action takes place, what changes occur in the field of vision?
	Speed	What is the rate of movement in the video?
Fragment Level	Frame Count	Could you please tell me how many frames I have inputted?
	Meaning of Order	In the sequence of frames provided, on which frame does the object first appear?
		In the sequence of frames provided, on which frame does the object last appear?
		In the sequence of frames provided, in which frames does the object exist?
	Frame Comparison	Are the two frames I provided exactly the same?
	Adjust or Not	These frames are all from the same video and capture the dynamic process of an action. The order of these frames may have been mixed up. Do we need to rearrange them to match the normal execution sequence of the action?
	Rearrangement	These frames are all from the same video and depict the dynamic process of an action. The order of these frames may have been mixed up. Based on the connections between the image frames, which of the following options represents the most appropriate sequence?

Table 2. Question templates authored by researchers undergo revision by GPT-4o, which rephrases them to maintain the original intent while introducing varied sentence structures and vocabulary.

Effect Type		Examples
Object Properties	Physical Properties	What modifications occur to the **wafer stick** as a result of the action? A. Not sure B. Nothing happened C. It broke D. It deformed
	Quantity	Once the action occurs, what changes are made to the **mugs**? A. There are about 5 or 6 mugs here B. There are about 1 or 2 mugs here C. There are about 3 or 4 mugs here D. Not sure
Object Relationships	Position	What adjustments take place in the **egg** following the action? A. An object appeared on top of it B. An object appeared in front of it C. An object appeared inside it D. An object appeared behind it
	Distance	What changes happen to the **chili** and the **cucumber** after the action is performed? A. They grew more distant B. It's unclear C. They came nearer D. Their separation remained consistent
	Similarity	What adjustments take place in the **box** following the action? A. One thing appeared above it B. Several things appeared above it, and they looked different from each other C. Not sure D. Several things appeared above it, and they looked similar to each other
Action Properties	Intensity	What alterations are observed in the **paper cups** after the action is taken? A. Not sure B. It collapsed C. It broke D. It remained standing
	Completion	After the action is done, what modifications occur to the **onion**? A. It appears unchanged from how it was initially B. Something was visible at the back of it C. An item appeared on its surface D. Something was detected below it
Special Actions	Slight Movement	What adjustments take place in the **shower pouf** during the action? A. I'm uncertain B. It dropped to the ground C. It was nearly at rest D. It ascended
	Mutiple-Object	What happens to the **two chargers** while the action is executed? A. They crossed paths B. They impacted each other C. They proceeded in the same direction D. It's unclear
	Compound	During the process of action, what modifications are observed in the **plate**? A. It fell after leaving the hand and did not come back B. It was continuously held without any separation C. It was detached from the hand but later reattached D. Unclear
Others	Camera movement	What alterations are evident in the **flower** while the action is carried out? A. It appeared to move to the right in view B. It appeared to ascend in view C. It appeared to move to the left in view D. I can't determine
	Surface Inclination	After the action is taken, what changes are noticed in the **cup**? A. It was stationary on a tilted surface B. It was stationary on a horizontal surface C. Not sure D. It rolled down a sloped surface

Table 3. Types of Effect Task

Effect Type (Random: 25.00)		LLaVA-NeXT-Video	MiniCPM-V 2.6	Video LLaMA 2.1	Qwen2-VL	ShareGPT4-Video	Video-CCAM	Avg.
Object Properties	Physical Properties	44.20	49.28	52.17	60.87	47.54	63.48	52.92
	Quantity	33.33	47.62	56.19	58.10	41.90	60.95	49.68
Object Relationships	Position	41.03	51.28	49.23	54.36	40.31	50.36	47.76
	Distance	39.56	46.67	40.89	40.44	40.44	48.44	42.74
	Similarity	42.86	49.52	47.62	52.38	38.10	59.05	48.25
Action Properties	Intensity	40.27	50.67	53.33	61.33	52.53	62.13	53.38
	Completion	39.31	43.68	38.85	35.63	48.05	34.02	39.92
Special Actions	Slight Movement	47.92	43.75	41.67	72.92	35.42	54.58	49.38
	Multiple-Object	50.00	60.67	76.67	66.67	40.67	58.67	58.89
	Compound	48.15	44.44	51.11	52.59	35.56	53.33	47.53
Others	Camera Movement	33.33	22.22	28.89	26.67	32.22	28.89	28.70
	Surface Inclination	28.57	49.52	58.57	60.48	41.43	51.43	48.33

Table 4. The results of the Effect task, dissected into more granular categories. Overall, Qwen2-VL achieved the best results, with Video-CCAM closely following. Notably, models exhibit suboptimal performance in distinguishing completed from incomplete actions, indicating a lack of ability to associate actions with the resulting state changes of objects.

Input		(Random)	LLaVA-NeXT-Video	MiniCPM-V 2.6	VideoLLaMA 2.1	Qwen2-VL	ShareGPT4Video	Video-CCAM
3	q1	25.00	20.33	93.82	42.86	97.25	60.99	14.18
	q2	25.00	19.23	48.90	35.71	29.12	76.15	38.35
	q3	33.33	46.96	80.66	71.27	71.82	88.41	66.34
	q4	33.33	69.23	65.38	81.54	80.00	75.55	80.06
	q5	25.00	23.85	23.08	33.08	27.69	23.68	23.36
4	q1	25.00	19.77	90.66	39.89	96.63	16.78	8.96
	q2	25.00	24.16	60.67	41.01	33.15	65.42	43.65
	q3	33.33	58.76	78.53	76.84	77.40	87.23	63.63
	q4	33.33	74.42	79.85	93.80	95.35	87.50	94.46
	q5	25.00	19.38	14.73	24.81	20.93	23.10	22.94
5	q1	25.00	17.98	86.44	7.45	96.05	0.00	47.61
	q2	25.00	28.81	59.89	50.28	37.85	41.00	55.24
	q3	33.33	55.68	67.61	80.11	74.43	89.69	64.83
	q4	33.33	82.81	84.38	94.53	96.88	91.55	96.49
	q5	25.00	18.75	16.41	22.66	18.75	23.29	23.92

Table 5. The results of all tasks in Fragment-Level under varying input frame counts. Questions q1 through q5 correspond to Frame Count, Meaning of Order, Frame Comparison, Adjust or Not, and Rearrangement, respectively.