VideoSSR: Video Self-Supervised Reinforcement Learning

Supplementary Material

1. Implementation Details

1.1. Evaluation Details for VideoSSR

1.1.1. Prompts

For Video QA tasks, we prompt the model to generate a direct answer. The specific prompt template utilized for these tasks is illustrated in Figure 1.

For Temporal Grounding tasks, our prompt format is based on the one utilized in the `lmms_eval` library [7], as depicted in Figure 2. While we observed that CharadesSTA seems to be particularly sensitive to prompt phrasing, we nonetheless applied this unified prompt across all benchmarks to ensure a fair and consistent evaluation.

For other specialized benchmarks, such as VCRBench, we adhere to the official prompts.

```
{question}\nAnswer with the option letter directly.
```

Figure 1. **Prompt template for Video QA tasks**

```
Please find the visual event described by a sentence in the video, determining its starting and ending times. The format should be: 'The event happens in the start time - end time'. For example, The event 'person turn a light on' happens in the 24.3 - 30.4 seonds. Now I will give you the textual sentence: {question} Please return its start time and end time.
```

Figure 2. **Prompt template for Temporal Grounding tasks**

1.1.2. Benchmarks

We adhered to specific evaluation protocols for several benchmarks to ensure fair and accurate assessment.

- **VinoGround:** We report the text score, which offers greater discriminative power between models.
- **Video-MME & LongVideoBench:** For both benchmarks, evaluations are conducted without the use of subtitles. For LongVideoBench, we specifically test on its validation set.
- **CGBench:** Our evaluation is performed on its 3k subset.
- **Temporal Grounding:** For benehmarks in this category, the model is required to predict a single most likely temporal interval. Results for QVHighlights and ActivityNet are reported on their validation sets.
- **VideoMMMU & Video-TT:** We report results on the multiple choice subset to facilitate answer extraction and comparison.
- **CVBench:** Our evaluation uses configurations of 32, 48, and 64 frames for each video, resulting in a significantly larger total number of frames processed per query.

1.1.3. Detailed Results

For Temporal Grounding tasks, we provide a more detailed breakdown of the results, as detailed in Table 1 and Table 2.

1.2. Evaluation Details for VIUBench

All evaluations on VIUBench utilized a fixed input of 48 frames with a maximum resolution of 512×512 pixels.

1.3. Dataset Composition and Statistics

We utilize Llava-Video [8] as the primary video source for constructing both VideoSSR-30K dataset and VIUBench. The detailed proportional distribution of data across the pretext tasks and their subtypes for both VIUBench and VideoSSR-30K is presented in Table 3.

2. Details of Pretext Tasks

2.1. Anomaly Grounding

Figure 3 illustrates the prompt template used for the Anomaly Grounding task. Table 4 provides the comprehensive list and definitions for all 14 perturbation subtypes designed for this task. The text in the "Description" column of the table is what replaces the `{description}` placeholder in the prompt for each respective subtype.

Notably, for perturbations targeting **Temporal Perception** (specifically *Slow* and *Fast*), we provided an expanded and highly detailed description within the prompt. This special note, as detailed at the bottom of Table 4, explicitly instructed the model to disregard the evenly spaced frame timestamps and instead rely solely on visual motion cues.

Despite this explicit guidance, the model's performance on these tasks remained notably poor. We hypothesize that this is because the base model, Qwen3-VL, has a strong inherent bias towards relying on textual timestamp information when it is available. Forcing the model to overcome this bias and learn true visual motion perception appears to be a significant challenge, even with detailed and explicit prompting.

Figures 6 through 9 illustrate several concrete examples of the Anomaly Grounding task, corresponding to four different perturbation types. For clarity, only a subset of key frames from each video is displayed. The model's objective

Table 1. **More results on QVHighlights and ActivityNet.**

Model	Frames	QVHighlights				ActivityNet			
		mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
Qwen3-VL-8B-Instruct [2]	32	43.7	62.3	42.5	24.2	36.5	52.3	34.5	18.3
	48	46.4	64.4	46.5	30.3	38.4	54.3	36.4	21.0
	64	48.6	64.5	48.6	33.9	39.8	55.6	38.6	23.0
VideoSSR-8B (Ours)	32	59.6(+15.9)	83.3(+21.0)	66.0(+23.5)	43.4(+19.2)	42.1(+5.6)	63.0(+10.7)	41.4(+6.9)	21.5(+3.2)
	48	61.1(+14.7)	83.5(+19.1)	66.9(+20.4)	48.3(+18.0)	43.0(+4.6)	63.2(+8.9)	42.3(+5.9)	22.7(+1.7)
	64	62.6(+14.0)	83.7(+19.2)	68.0(+19.4)	49.7(+15.8)	43.7(+3.9)	63.3(+7.7)	42.7(+4.1)	24.2(+1.2)

Table 2. **More results on CharadesSTA and Tacos.**

Model	Frames	CharadesSTA				Tacos			
		mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7
Qwen3-VL-8B-Instruct [2]	32	50.3	76.5	58.1	27.9	22.4	34.7	19.2	7.1
	48	50.0	76.6	56.1	26.9	25.9	39.0	24.0	10.7
	64	49.2	77.1	54.2	25.5	28.1	42.0	26.6	12.3
VideoSSR-8B (Ours)	32	52.1(+1.8)	78.2(+1.7)	60.6(+2.5)	30.8(+2.9)	23.1(+0.7)	34.1(-0.6)	19.8(+0.6)	7.4(+0.3)
	48	51.1(+1.1)	79.0(+2.4)	59.9(+3.8)	27.5(+0.6)	27.7(+1.8)	40.0(+1.0)	24.7(+0.7)	12.3(+1.6)
	64	49.9(+0.7)	78.7(+1.6)	57.6(+3.4)	24.2(-1.3)	30.6(+2.5)	43.8(+1.8)	28.1(+1.5)	14.4(+2.1)

Table 3. **Task Distribution in VIUBench and VideoSSR-30K.**

Task	Subtype	VIUBench	VideoSSR-30K
Grounding	*Total*	**1500**	**12000**
	Shuffle	300	2400
	Mirror	300	2400
	ZoomOut	300	2400
	Rotata	300	2400
	Channel	300	2400
Jigsaw	*Total*	**600**	**12000**
	Easy	300	6000
	Hard	300	6000
Counting	*Total*	**600**	**6000**
	Easy	300	3000
	Hard	300	3000
TOTAL		**2700**	**30000**

In a segment of this video, {description}. Your task is to identify the precise time interval of this change. Please only provide the start and end times in seconds, formatted as <start_time>-<end_time> (e.g., '14.5-26.2').

Figure 3. **Prompt template for Anomaly Grounding.**

is to predict the temporal range of the introduced anomaly based on the visual evidence.

2.2. Object Counting

Figure 4 illustrates the prompt template used for the Object Counting task. Concrete visual examples of this task are provided in Figure 10 and Figure 11.

Count the number of circles, squares, and triangles that appear in this video. Be aware that the shapes can appear in any color and at any angle of rotation. They may be present on one or multiple frames, and any given frame can contain more than one shape. Provide the answer as three comma-separated numbers in the format: circles,squares,triangles. For example, if you see 3 circles, 1 square, and 4 triangles, your answer should be '3,1,4'.

Figure 4. **Prompt template for Object Counting.**

2.3. Temporal Jigsaw

Figure 5 shows the prompt template for the Temporal Jigsaw task. Figure 12 provides a concrete visual example of the shuffled video sequence that is presented to the model. For a clearer understanding of the task and to provide a direct comparison, the corresponding original video with the clips in their correct temporal order is also shown in Figure 13.

Table 4. **Definitions of the 14 Perturbation Subtypes for Anomaly Grounding.** For temporal perception tasks, an additional detailed note (marked with *) was provided to guide the model.

Category	Perturbation Type	Description
Fine-Grained Perception		
	Saturation	the colors in the video become oversaturated and unnaturally vibrant.
	Noise	Gaussian noise is added to the video.
	Blur	the video becomes blurry or out of focus.
	Grayscale	the video becomes black and white.
	Invert	the colors in the video are inverted.
	Channel Swap	the red and blue color channels in the video are swapped.
Spatial Perception		
	Zoom In	the video is zoomed in.
	Rotate	the video is rotated 180 degrees.
	Zoom Out	the video is zoomed out.
	Mirror	The video is mirrored horizontally.
Temporal Perception		
	Slow	the video slows down, this means the action unfolds at an unusually slow pace, making movements appear prolonged.*
	Fast	the video speeds up, this means the segment plays at a high speed, compressing the action and making movements appear jerky or rushed.*
	StutterHold	the video appears to freeze and stutter on a few frames, this means instead of playing smoothly, the video repeatedly freezes on a single frame before jumping to the next.
	Shuffle	the frames are shuffled, this means the order of events is scrambled, making the action appear illogical and chaotic.

***Special Note for Slow/Fast perturbations:** To ensure a fair challenge, even if the video's actual speed changes (e.g., slow motion or fast forward), the timestamps for each frame have been intentionally kept evenly spaced. This creates the illusion of a constant playback speed. Therefore, you should not rely on the timestamps when judging the speed. Instead, your judgment must be based solely on the visual content. You should analyze the motion within the video itself by observing how much or how little the scene changes between consecutive frames to determine the true playback speed.

This video is presented as 6 separate clips, which are in a shuffled order. Your task is to determine the correct chronological sequence. Please output a six-digit number that specifies the order in which to play the clips you are seeing (labeled 1 through 6 by their position). For example, if you decide that the clip at position 3 of this video is the true beginning (the 1st clip of the original video), the clip at position 4 is the 2nd part, the clip at position 1 is the 3rd part, the clip at position 5 is the 4th part, the clip at position 6 is the 5th part, and finally the clip at position 2 is the 6th part, then your answer should be '341562'.

Figure 5. **Prompt template for Temporal Jigsaw.**

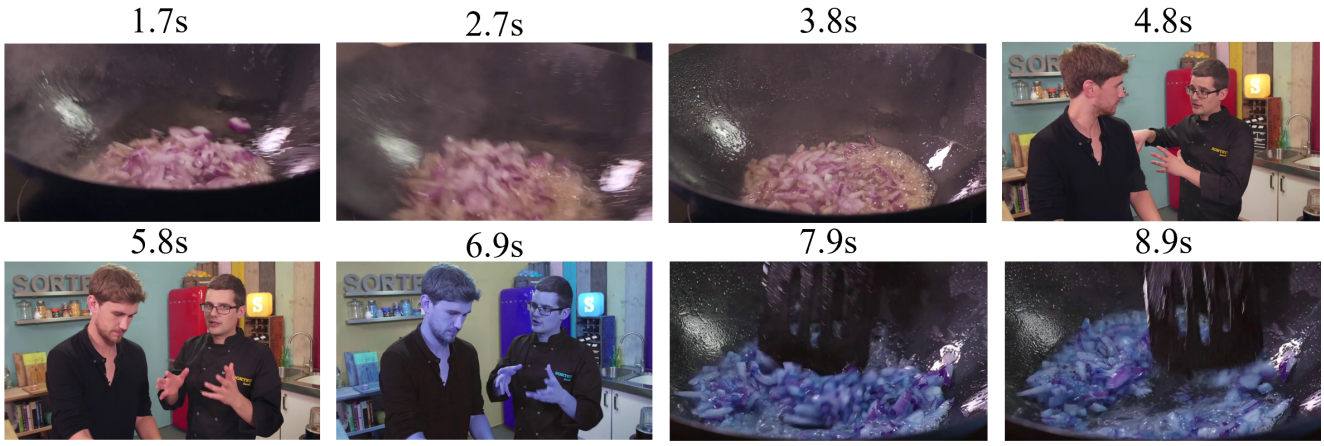

Figure 6. **An example of Channel Swap.** The ground truth is 6.9s–9.2s.

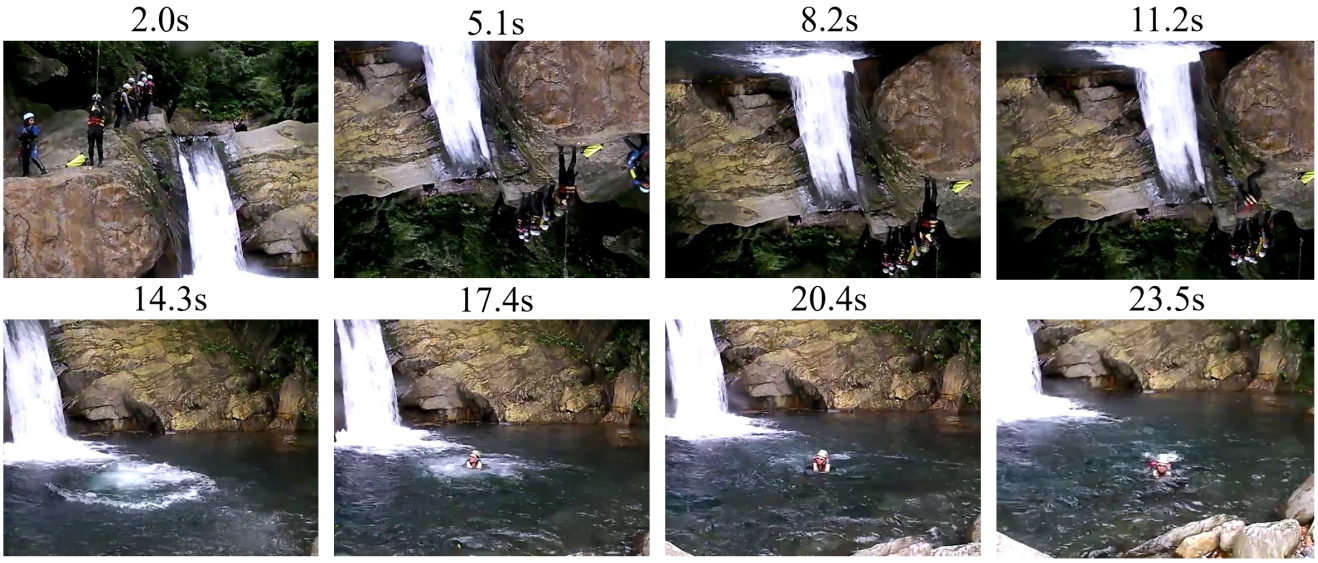

Figure 7. **An example of Rotate.** The ground truth is 5.1s–11.7s.

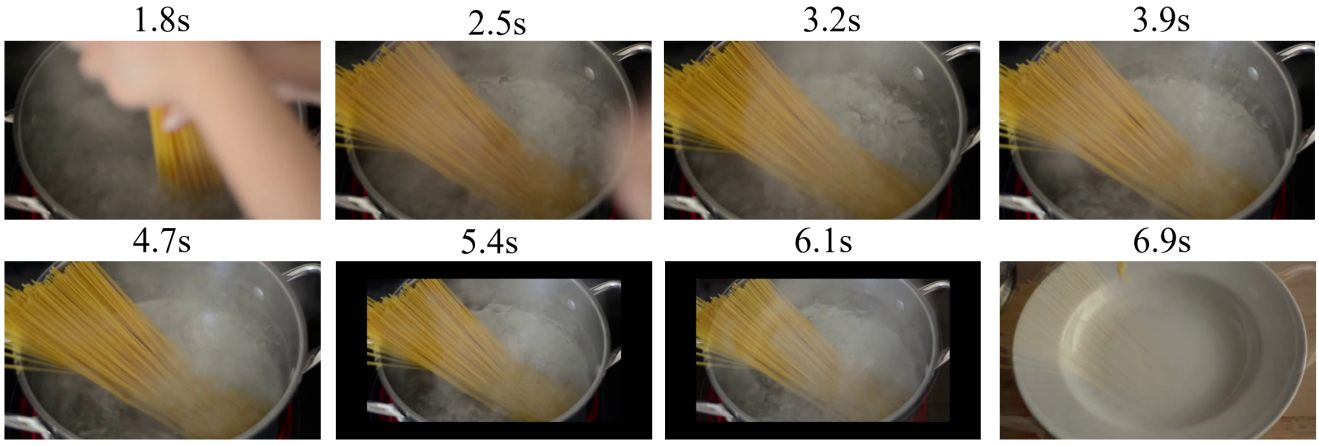

Figure 8. **An example of ZoomOut.** The ground truth is 5.4s–6.9s.

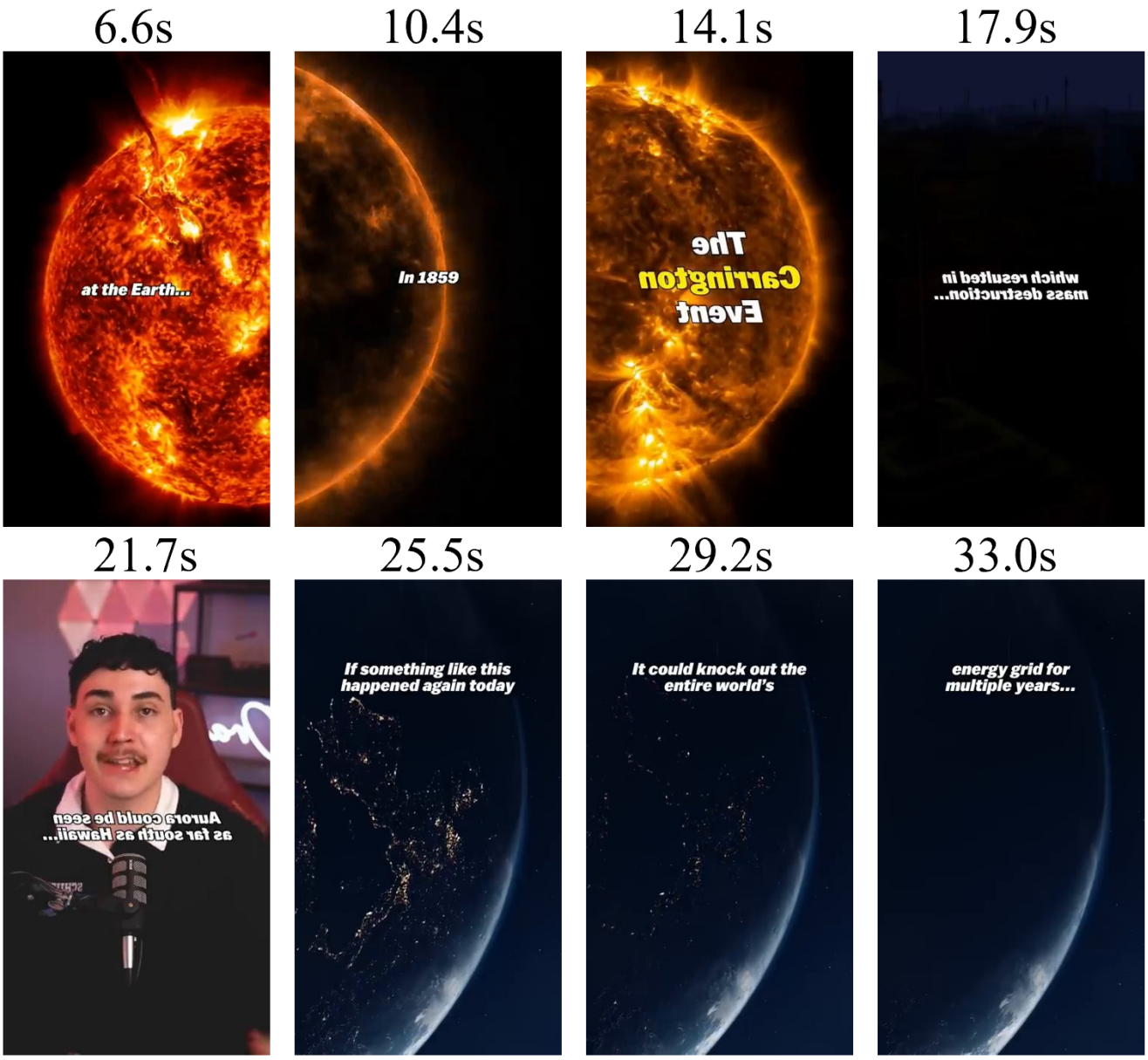

Figure 9. **An example of Mirror.** The ground truth is 14.1s–22.6s.

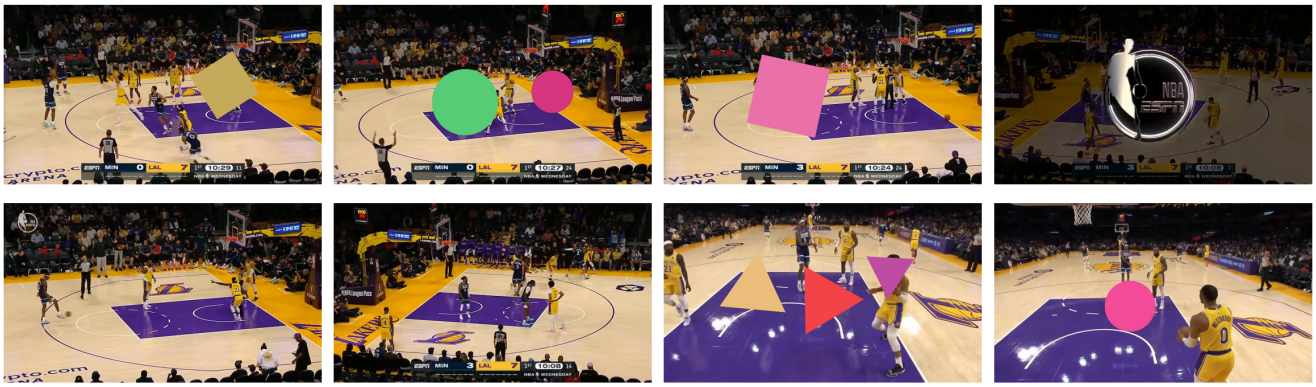

Figure 10. **An example of Object Counting.** The ground truth (circles, squares, and triangles) is 3,2,3.

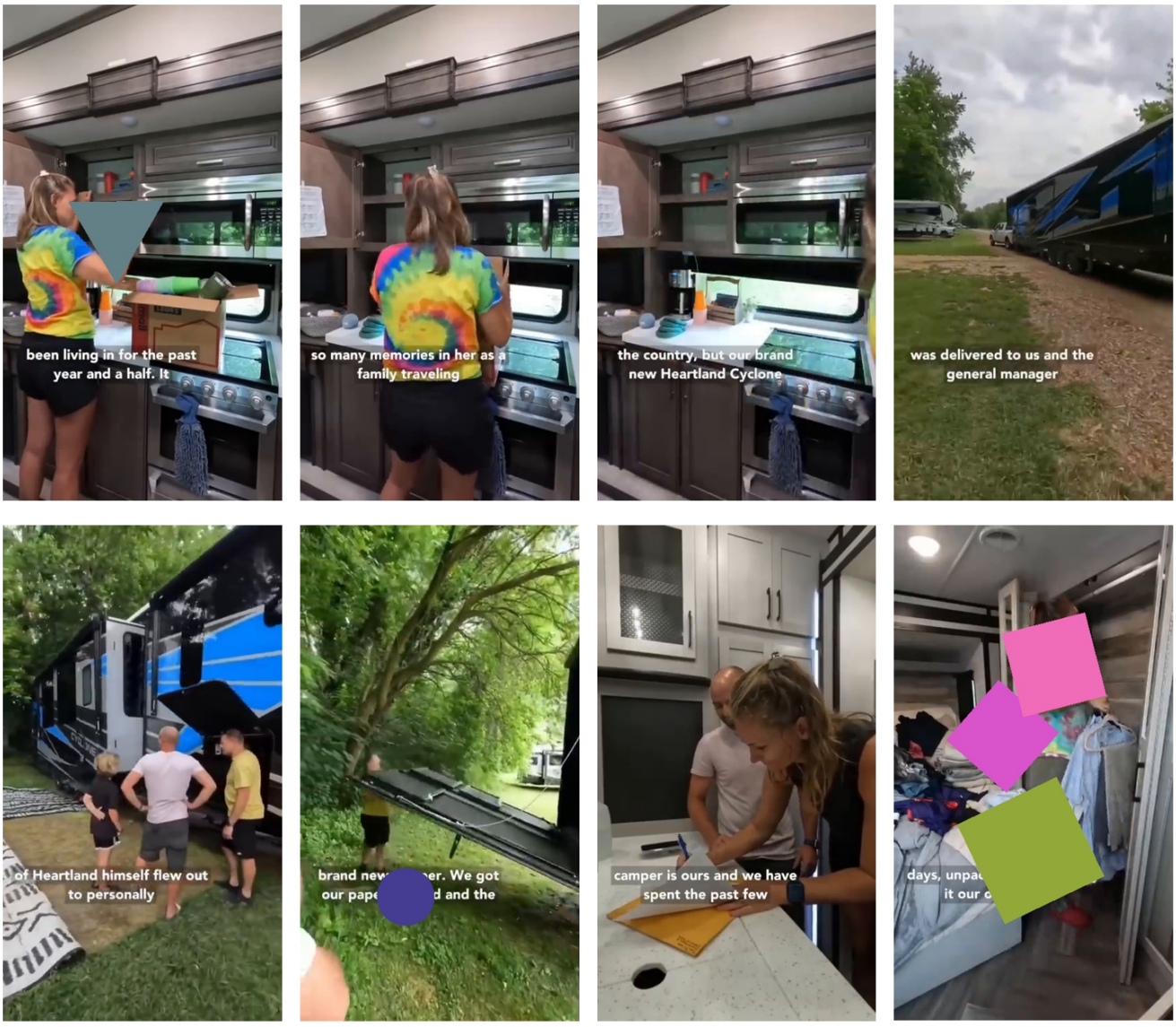

Figure 11. **An example of Object Counting.** The ground truth (circles, squares, and triangles) is 1,3,1.

Figure 12. **An example of Temporal Jigsaw.** The ground truth is 452316. The corresponding unshuffled video is shown in Figure 13.

Figure 13. **The original video corresponding to the example in Figure 12.**

2.4. Exploration of Alternative Pretext Tasks

In addition to the three pretext tasks detailed in the main paper, we also investigated other self-supervised learning paradigms. Our exploration included generative modeling approaches, such as masked [1, 3] frame reconstruction and autoregressive [4, 6] next frame prediction. Furthermore, we experimented with a task focused on direct temporal speed prediction [5].

However, our preliminary experiments indicated that these alternative tasks did not yield significant or consistent performance improvements on our downstream evaluation benchmarks. This suggests that while these methods are powerful, their objectives may not be as directly aligned with cultivating the high level perceptual and reasoning skills targeted by our final task selection. The discovery of an even broader range of effective self-supervised tasks for enhancing MLLMs remains a promising direction for future work.

References

[1] Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Bohan Zeng, Yang Shi, Sihan Yang, Pengfei Wan, Qiang Liu, Liang Wang, and Tieniu Tan. Versavid-r1: A versatile video understanding and reasoning model from question answering to captioning tasks. *arXiv preprint arXiv:2506.09079*, 2025. 9

[2] Qwen. Qwen3-vl, 2025. 2

[3] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 9

[4] Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction. *arXiv preprint arXiv:2505.22457*, 2025. 9

[5] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pages 504–521. Springer, 2020. 9

[6] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Danhao Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025. 9

[7] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 1

[8] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1