

MASH-VLM: Mitigating Action-Scene Hallucination in Video-LLMs through Disentangled Spatial-Temporal Representations

Supplementary Material

In this supplementary material, we present implementation details, additional ablation studies, further qualitative evaluations, and examples from the UNSCENE benchmark to complement the main paper. The supplementary material is organized as follows:

7. Implementation details
8. Additional ablation studies
9. Further qualitative evaluations
10. Examples from the UNSCENE benchmark

1. Implementation details

In this study, we use CLIP-ViT-Large-Patch14-336 [6] as a vision encoder and a two-layer MLP as a projector. The visual encoder, projector, and LLM weights are initialized with the pre-trained weights of LLaVA v1.5 [3], which employs Vicuna-v1.5 [7] with 7B parameters as the language model. During instruction tuning, we freeze the vision encoder, allowing only the projector and the LLM to be fully fine-tuned. Following prior work [3], we set the learning rate to $2e^{-5}$, the total batch size to 128, and train the model for 2 epochs. We adopt a cosine decay learning rate schedule with a warmup ratio of 0.03, AdamW [5] as the optimizer with no weight decay, and DeepSpeed Stage 3.

2. Additional ablation study

Ablation study on spatial and temporal features for video. In Table 1, we conduct an ablation study on various video features. When using both temporal and spatial features during training, our MASH-VLM achieves an accuracy of 41.81% on UNSCENE Binary and 53.41% on MVBench, demonstrating a significant improvement over using only one of these features. Furthermore, incorporating CLS tokens and frame-difference tokens into the temporal features enhances performance, highlighting the effectiveness of our proposed feature extraction method.

Ablation study on LLM Tuning Scheme. Table 2 presents the performance results across different LLM tuning schemes. Full fine-tuning achieves the best performance, with accuracy of 57.85% on UNSCENE Binary and

Table 1. **Ablation study on spatial and temporal features for video.** SP, TP, and F-Diff refer to spatial pooling, temporal pooling, and frame difference, respectively.

Temporal token	Spatial tokens	UNSCENE Binary	MVBench
\times	TP	9.94	37.15
SP	\times	28.19	45.98
SP + CLS	\times	30.30	48.02
SP + CLS + F-Diff	\times	31.29	48.77
SP + CLS + F-Diff	TP	41.81	53.41

Table 2. **Ablation study on LLM Tuning Scheme.**

LLM Tuning	UNSCENE Binary	MVBench
Frozen	37.49	48.41
LoRA	44.31	51.10
Full F.T.	57.85	57.78

57.78% on MVBench, outperforming both freezing LLM parameters and using LoRA [1] tuning.

Effect of token types during inference. In Table 3, we investigate whether the spatial and temporal tokens of MASH-VLM preserve their respective information in a disentangled manner. When only spatial tokens are used during inference, the performance on action-related QA for unusual context videos, which require temporal understanding, drops by 11.24 points. Conversely, using only temporal tokens during inference results in a performance decrease of 23.64 points on scene-related QA for scene-only videos and 18.1 points for unusual context videos, both of which require spatial understanding. These results demonstrate that the disentangled tokens effectively preserve their respective information: spatial tokens retain spatial details, while temporal tokens capture temporal dynamics. Furthermore, by leveraging both disentangled tokens, MASH-VLM achieves the highest performance.

Table 3. Effect of token types during inference.

Video token types	Scene-only	Unusual context	
	Scene	Action	Scene
Spatial	49.75	27.15	74.85
Temporal	31.27	33.13	62.15
Spatial & Temporal	54.91	38.39	80.25

3. Further qualitative evaluations

Qualitative results. In Figure 1, we present qualitative comparisons with other methods. In the top example, a scene-only video depicting an ice hockey rink without any people is shown. Previous video-LLMs incorrectly respond that people are present in the background or that a team is playing a game. In contrast, MASH-VLM not only accurately predicts the absence of people but also provides a detailed description of the background. In the middle example, an unusual context video shows a person performing parkour in a restroom. While VideoChat2 [2] incorrectly predicts that no one is performing parkour, both ST-LLM [4] and MASH-VLM correctly identify the presence of a person engaged in parkour. Similarly, as shown in the bottom example, MASH-VLM accurately identifies the background. These qualitative results demonstrate that MASH-VLM effectively mitigates action-scene hallucinations.

Attention scores. We compare the attention scores of the baseline model (without DST-attention and Harmonic-RoPE) and MASH-VLM as shown in Figure 2. (a) The baseline model shows improper focus: when answering about actions, the model tends to focus on spatial tokens (left), and when answering about scenes, it attends to both spatial and temporal tokens (right), resulting in action-scene hallucination. (c) When answering about actions, the baseline model tends to focus on spatial tokens (left), and when answering about scenes, it attends to spatial tokens but generates a hallucinated response (right). This hallucination arises from the baseline model’s failure to disentangle spatial and temporal tokens, leading to entanglement between spatial and temporal tokens. (b,d) In contrast, MASH-VLM appropriately focuses on temporal or spatial tokens depending on the question type and learns disentangled spatial and temporal representation, effectively mitigating action-scene hallucination.

4. Examples of UNSCENE benchmark

In Figure 3 and Figure 4, we showcase examples of unusual context videos in the UNSCENE benchmark. We also present examples of scene-only videos as shown in Figure 5. We also provide example videos of UNSCENE benchmark in the supplementary material.

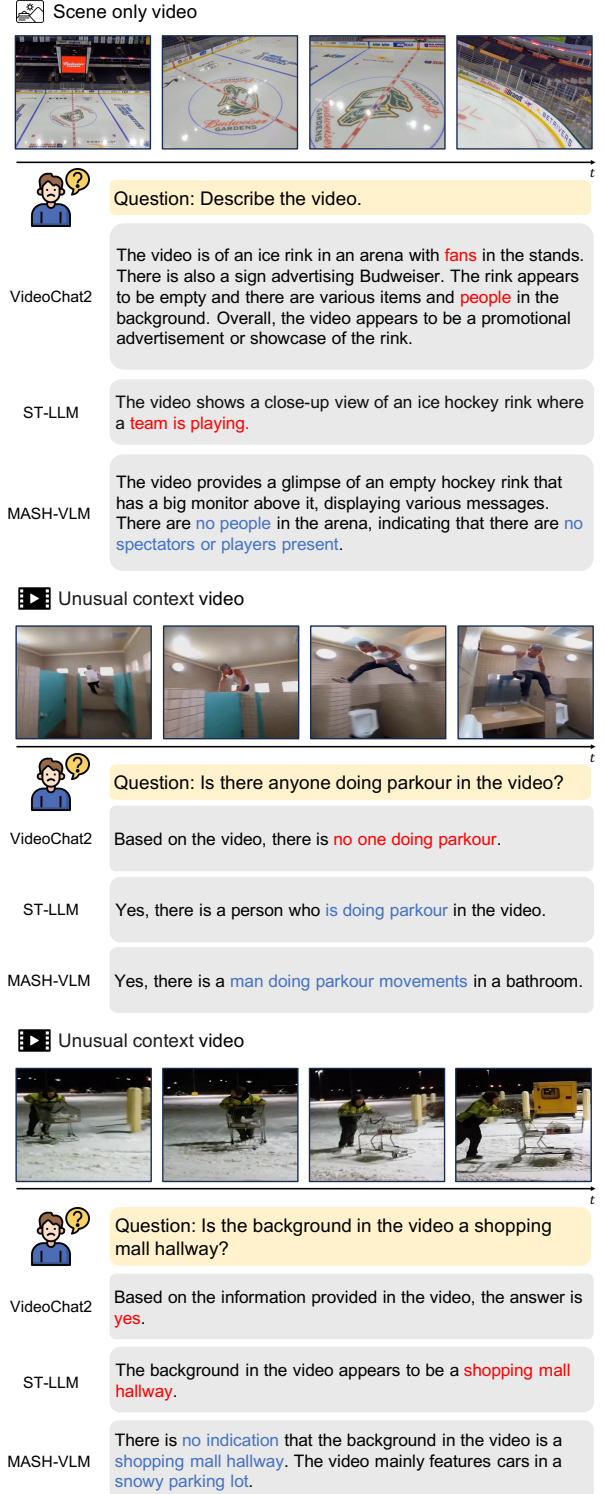


Figure 1. Qualitative results on the UNSCENE benchmark.

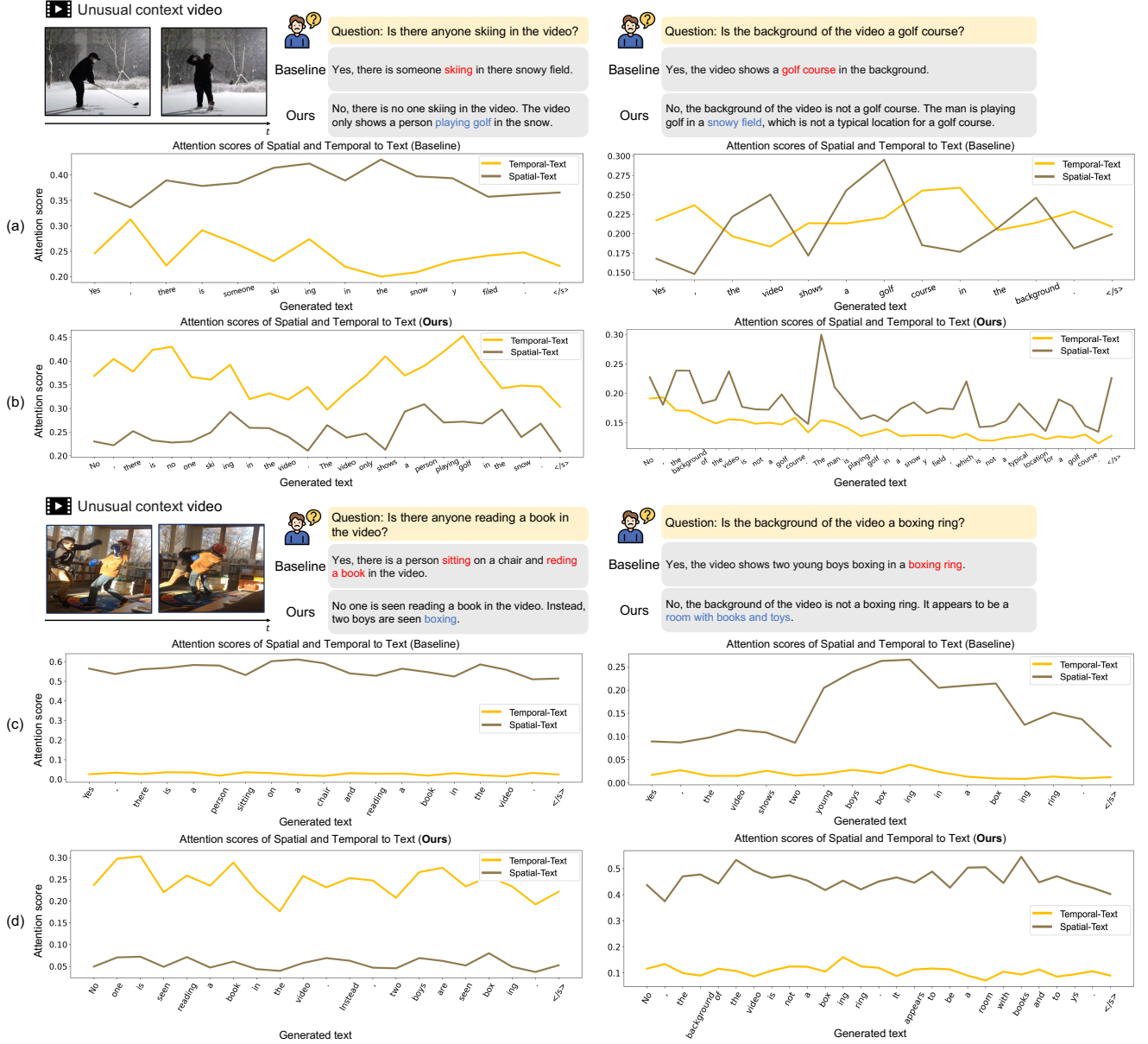


Figure 2. **Attention scores of spatial-to-text and temporal-to-text token attention.** We compare the attention scores of the baseline model (without DST-attention and Harmonic-RoPE) and MASH-VLM. (a) The baseline model shows improper focus: when answering about actions, the model tends to focus on spatial tokens (left), and when answering about scenes, it attends to both spatial and temporal tokens (right), resulting in action-scene hallucination. (b) MASH-VLM appropriately focuses on temporal or spatial tokens depending on the question type. (c) When answering about actions, the baseline model tends to focus on spatial tokens (left), and when answering about scenes, it attends to spatial tokens but generates a hallucinated response (right). This hallucination arises from the baseline model’s failure to disentangle spatial and temporal tokens, leading to entanglement between spatial and temporal tokens. (d) In contrast, MASH-VLM not only focuses on temporal or spatial tokens depending on the question type but also learns disentangled spatial and temporal representation, effectively mitigating action-scene hallucination.

UNSCENE: UNusual context & SCENE only benchmark

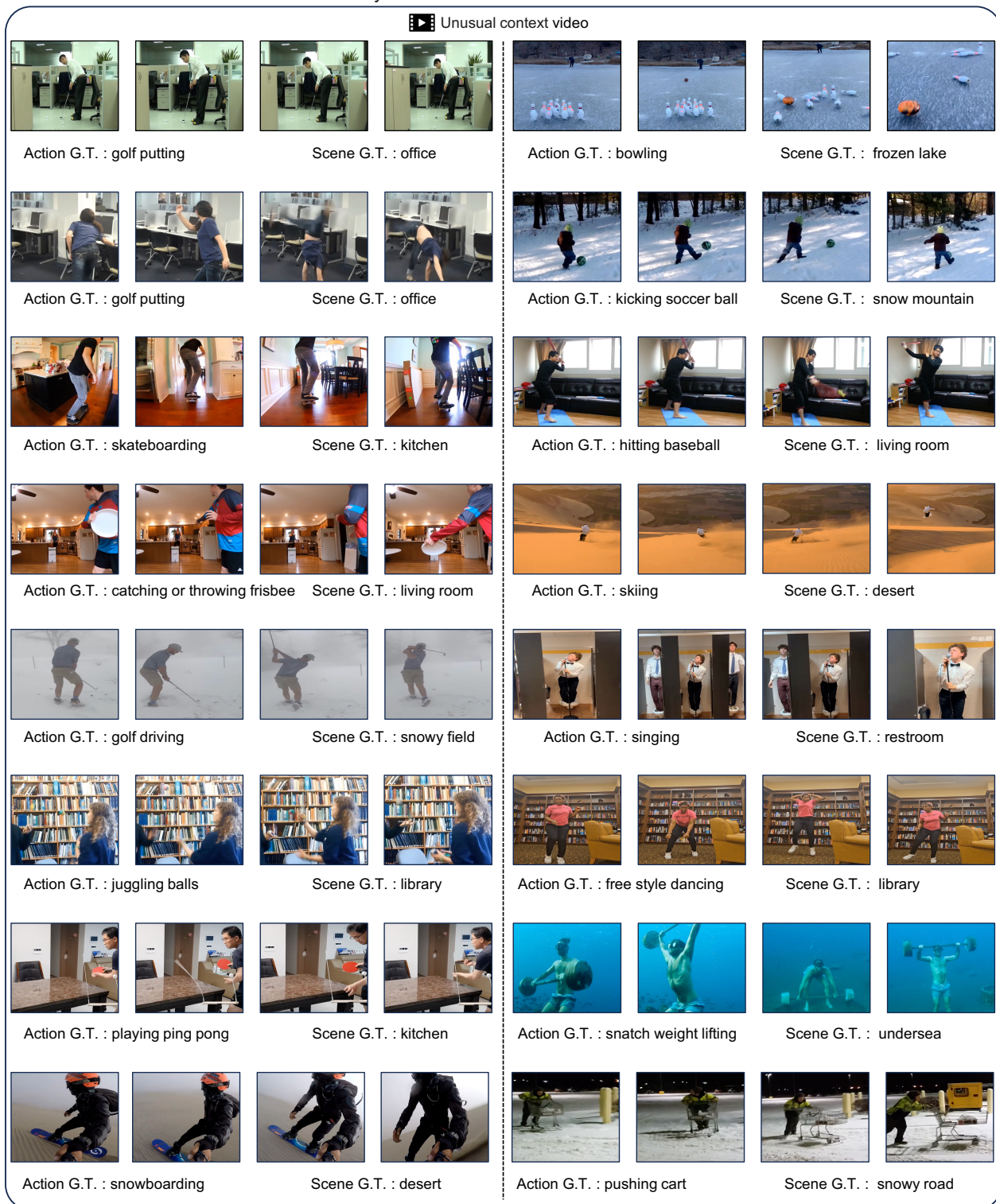


Figure 3. Examples of unusual context videos in the UNSCENE benchmark.

UNSCENE: UNusual context & SCENE only benchmark

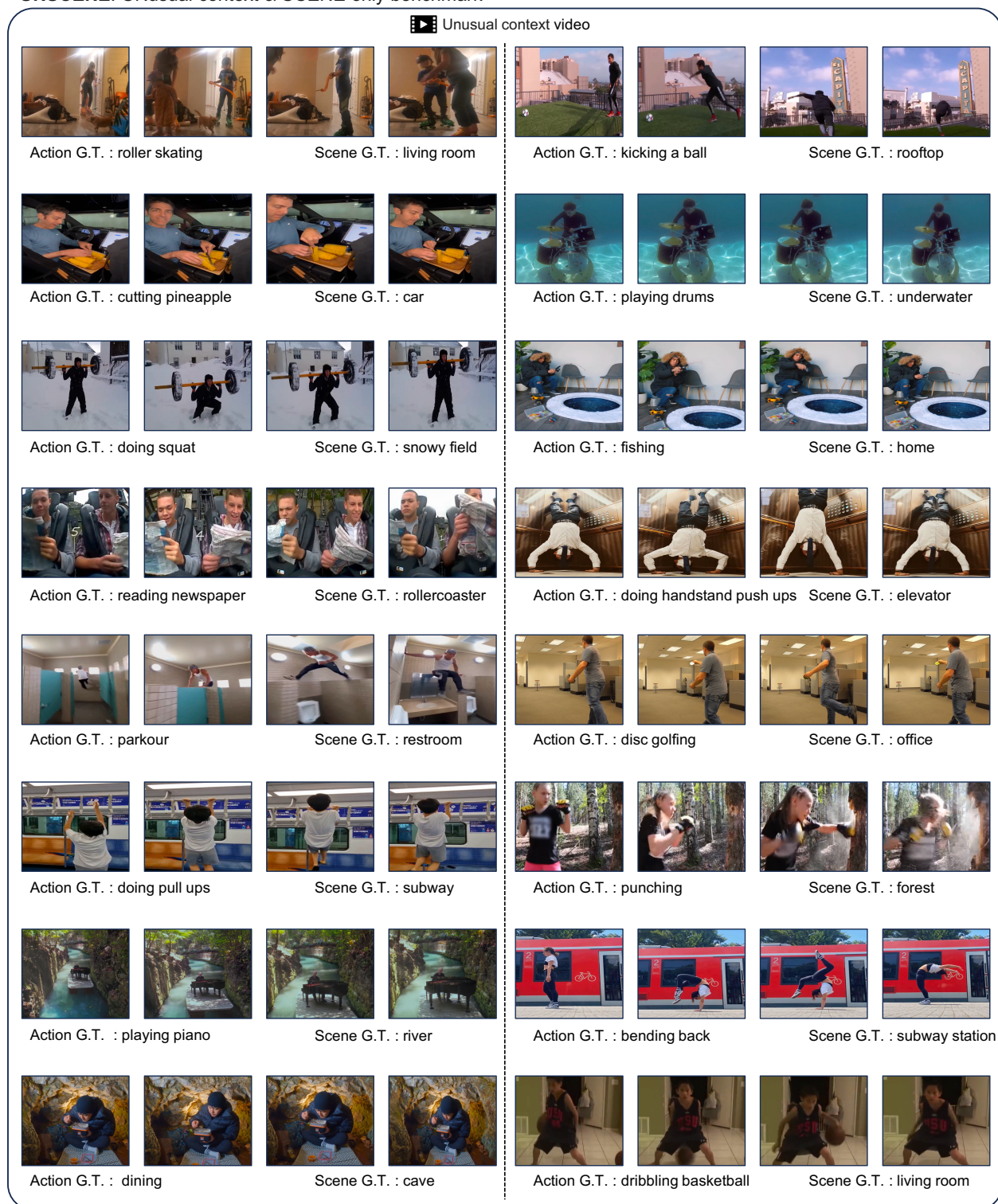

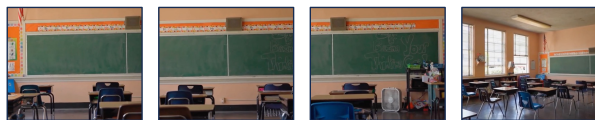


Figure 4. Examples of unusual context videos in the UNSCENE benchmark.

UNSCENE: UNusual context & SCENE only benchmark

 Scene only video



Scene G.T. : classroom



Scene G.T. : office



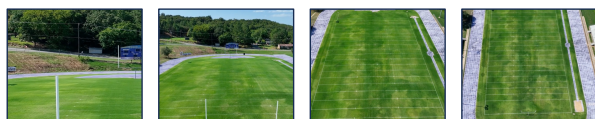
Scene G.T. : billiards hall



Scene G.T. : restroom



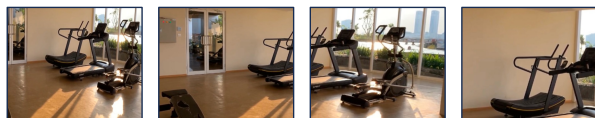
Scene G.T. : theater



Scene G.T. : football field



Scene G.T. : swimming pool



Scene G.T. : gym



Scene G.T. : office



Scene G.T. : hockey rink



Scene G.T. : kitchen



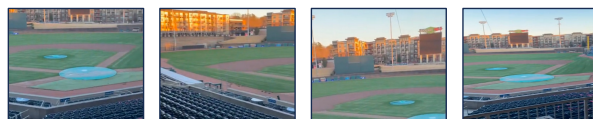
Scene G.T. : football field



Scene G.T. : basketball court



Scene G.T. : tennis court



Scene G.T. : baseball stadium



Scene G.T. : escalator

Figure 5. Examples of scene only videos in the UNSCENE benchmark.

References

- [1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. [1](#)
- [2] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. [2](#)
- [3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. [1](#)
- [4] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *ECCV*, 2024. [2](#)
- [5] I Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. [1](#)
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. [1](#)
- [7] The Vicuna Team. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. <https://lmsys.org/blog/2023-03-30-vicuna/>. [1](#)