VideoRefer Suite: Advancing Spatial-Temporal Object Understanding with Video LLM
- Supplementary Material -

Yuqian Yuan[1,2,*], Hang Zhang[2], Wentong Li[1], Zesen Cheng[2], Boqiang Zhang[2], Long Li[1,2,*]
Xin Li[2], Deli Zhao[2], Wenqiao Zhang[1,†], Yueting Zhuang[1], Jianke Zhu[1,†], Lidong Bing[3]
[1]Zhejiang University [2]DAMO Academy, Alibaba Group [3]Shanda AI Research Institute

The supplementary material is organized as follows:
- § 1: More background;
- § 2: More experimental results;
- § 3: Additional implemental details;
- § 4: More details of VideoRefer-700K and Benchmark;
- § 5: Limitations.

1. More Background

1.1. Video Large Language Models

Large Language Models (LLMs) have revolutionized the field of artificial intelligence by proving their capability to tackle diverse tasks related to language comprehension and generation. To fully leverage the potential of LLMs for visual understanding, researchers have increasingly turned their attention to image-based Multimodal Large Language Models (MLLMs) [1, 5, 11, 14–17, 19, 36], which integrate language and visual data within a unified feature space. This integration has emerged as a significant area of research focus. In parallel, Video Large Language Models (Video LLMs) [6, 9, 10, 12, 12, 13, 18, 18, 22, 32, 34, 37] have garnered increasing attention fueled by advancements in image-based MLLMs. Most Video LLMs primarily follow the trend of utilizing pre-trained vision models to extract sequence-based information from videos, which is then interleaved with textual embeddings for LLM to generate responses [23]. Despite their promising results, current Video LLMs still face challenges in fine-grained regional and temporal understanding.

1.2. Regional Understanding with MLLMs

To attain fine-grained regional object-level comprehension, MLLMs can be incorporated with instance-level visual representations. This integration allows models to generate semantic understandings that focus on specific regions. In the context of image-based MLLMs, recent researchs [2–4, 7, 8, 21, 24, 26, 27, 29–31, 33, 35, 38] has demonstrated a significant trend to enable the image referring with spatial visual prompts. In contrast, research focused on video-based regional understanding across dynamic sequence-based scenes is relatively limited. Merlin [28] first explored video-based referring and future reasoning by employing three manually selected frames as visual input, which limits the model's ability to comprehend longer and more intricate scenes. Elysium [25] introduces a million-scale dataset for object-level tasks in videos; however, the provided descriptions tend to be quite simplistic. Another reseach work is Artemis [20], but it primarily emphasizes basic single object descriptions, thereby constraining its capacity to analyze relationships among various objects or perform more complex tasks on specific objects within dynamic sequences. Moreover, Artemis utilizes a sparse bounding box representation, which inadequately captures the nuances of the objects. Compounding these challenges is the lack of large-scale, high-quality region-level video instruction data and benchmarks for thorough evaluation, which further hampers progress in this domain. To address these issues, we introduce the VideoRefer Suite to advance spatial-temporal understanding.

2. More Experimental Results

2.1. Additional Ablation Studies

Ablation on VideoRefer-700K Dataset. Table A1 summarizes the ablation results for various data types in the constructed VideoRefer-700K dataset. The results indicate that using a short description yields a score of 2.43 on Bench^{D} and 68.3 on Bench^{Q}, along with an MVBench score of 58.0. Incorporating question-answering (QA) data improves the performance to 2.45 for Bench^{D} and 71.7 for Bench^{Q}, while maintaining an MVBench score of 58.4. Notably, the method employing detailed descriptions achieves the best results, with scores of 3.42 on Bench^{D}, 71.9 on

*Work is done during internship at DAMO Academy, Alibaba Group
†Corresponding author

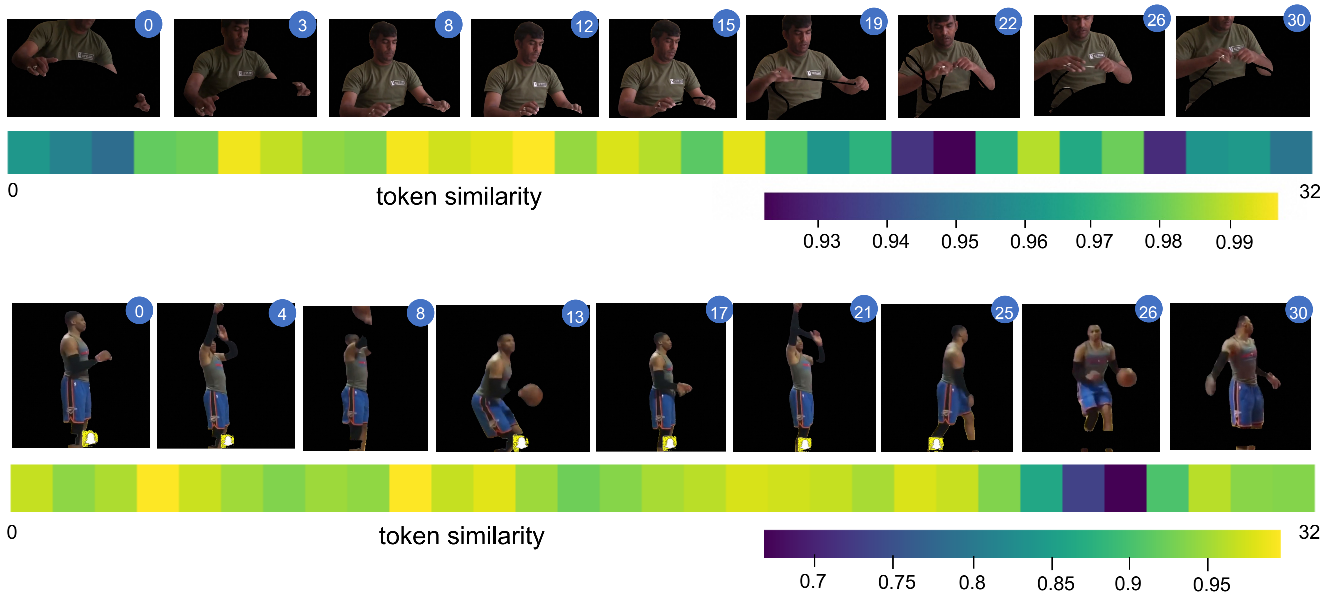


Figure A1. Visualizations of similarity among adjacent object-level token pairs across the temporal dimension. Here, we use cosine similarity as the measurement.

	Method	Bench^{D}	Bench^{Q}	MVBench
0	w/o Regional data	–	–	57.9
1	+ Short description	2.43	68.3	58.0
2	+ QA	2.45	71.7	58.4
3	+ Detailed description	**3.42**	**71.9**	**59.6**

Table A1. Ablation results on various data types in VideoRefer-700K dataset. Bench denotes VideoRefer-Bench for simplicity.

Model	MVBench	VideoMME	$\text{VideoRefer}^{\text{D}}$	$\text{VideoRefer}^{\text{Q}}$
Three Stages	57.2	55.5/57.5	**3.44**	70.3
Four Stages	**59.6**	**55.9/57.6**	3.42	**71.9**

Table A2. Comparisons with different training stages.

Bench^{Q}, and 59.6 on MVBench. These results demonstrate that the inclusion of more comprehensive data significantly enhances overall performance.

Ablation on Training Stage Paradigm. We conduct experiments involving a four-stage training process. The first two stages focus on scene-level and object-level alignment, respectively. Stages 2.5&3 have identical training setups but differ in data, with Stage 2.5 using descriptive data with uniform prompts. To prevent such a large amount of data from impacting the model's instruction-following ability, we separate these stages to allow initial acquisition of detailed knowledge before SFT. Additionally, we evaluated the impact of merging the final two stages, as shown in Table A2. The results reveal that a four-stage training approach significantly enhances instruction-following ability, evidenced by considerable improvements in MVBench (+2.4) and $\text{VideoRefer}^{\text{Q}}$ (+1.6).

Union u	$\text{VideoRefer-Bench}^{\text{D}}$		$\text{VideoRefer-Bench}^{\text{Q}}$	
	TD	HD	SQ	RQ
32	3.17	3.01	68.7	58.1
16	**3.20**	2.99	69.3	58.5
8	3.18	3.02	69.6	57.8
4	3.10	**3.04**	**70.6**	60.5
1	3.08	2.98	68.9	**60.9**

Table A3. Temporal and sequential performance comparisons for various union u in the TTM module under multi-frame mode.

Impacts of Different Union Numbers in TTM. The Temporal Token Merge (TTM) Module is designed to capture essential object-level tokens across the temporal dimension in multi-frame mode. Fig. A1 visualizes the similarity scores between adjacent object token pairs. It is evident that most adjacent tokens exhibit high similarity, making it necessary to merge those tokens with significant similarity. We conducted ablation experiments using temporal and sequential metrics to investigate the effects of varying numbers of token unions u. The experimental results are detailed in Table A3. Notably, with $u = 4$, VideoRefer achieves the best performance in Hallucination Detection (HD) and Sequential Questons (SQ), and ranks second in Reasoning Questions (RQ). We adopt $u = 4$ to strike a balance between performance and token costs in our approach.

Different Reference Forms. Our model is capable of supporting various types of visual input, including points, boxes, and masks. However, as shown in Table A4, the

Prompts	VideoRefer^{D}	VideoRefer^{Q}
Point	3.15	67.8
Box	3.30	70.3
Mask	**3.42**	**71.9**

Table A4. Different visual prompts.

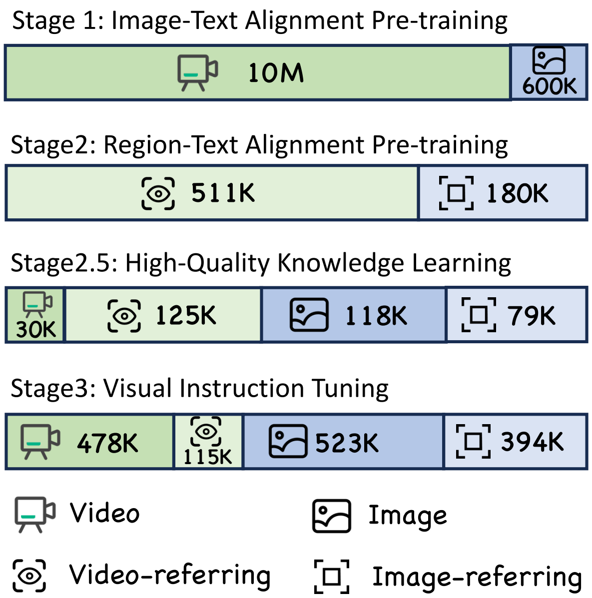


Figure A2. Visual illustrations of the data distribution for each training stage.

features encoded by masks tend to be more precise, leading to better performance, whereas points are less precise by comparison.

2.2. More Qualitative Results

We provide additional visualization results to emphasize performance across a variety of tasks, such as single-object referring, video relationship analysis, complex reasoning, future prediction, and video object retrieval. Besides, we present the examplar cases to demonstrate the capabilities in general video understanding and image object understanding. Fig. A6 showcases these visual examples. A randomly selected mask along with its corresponding frame is used as the region input.

3. Additional Implemental Details

3.1. Training Stages

The training pipeline of our model is structured into four distinct stages. Fig. A2 presents the data distribution for each stage.

Stage 1: Image-Text Alignment Pre-training. In this initial pre-training phase, we utilize the same dataset as employed in the first stage of VideoLLaMA2.1 [6]. During this phase, the parameters of both the vision encoder and the large language model are frozen, and training is conducted solely on the STC connector [6], enabling the alignment of image and text modalities.

Stage 2: Region-Text Alignment Pre-training. This stage further incorporates the Object Encoder to capture object-level features based on the weights obtained from Stage 1. The training focus is exclusively on the spatial-temporal Object Encoder to ensure the alignment of intricate object-level features with corresponding language embeddings. We use the generated 500K region-level short descriptions, along with video and image referring segmentation datasets as the training data. During this stage, all the data are processed in single-frame mode to focus solely on alignment.

Stage 2.5: High-Quality Knowledge Learning. At this intermediate stage, the weights of vision encoder remain frozen, while the STC connector, Object Encoder, and LLM undergo fine-tuning. This stage aims to infuse the model with high-quality captioning data, utilizing a diverse dataset that includes 118K image-caption pairs, 30K video-caption pairs, 79K image-level region caption data, and 125K video-level region caption data, inclusive of the detailed descriptions we curated. For object-level video data, we employ a balanced approach, using half in single-frame mode and half in multi-frame mode.

Stage 3: Visual Instruction Tuning. The training configuration for this stage closely mirrors that of Stage 2.5. The primary objective is to enhance the model's ability to accurately interpret user instructions and tackle complex object-level understanding tasks. For video-level data, we utilize the same dataset segments as those used in VideoLLaMA2.1 [6]. For image-level data, we employ the datasets from LLaVA [16]. In addition, we incorporate 294K image-level region data and 115K previously constructed video-level region data to further strengthen the model's capabilities. We also employ a balanced approach using half in single-frame mode and half in multi-frame mode in this stage.

4. More Details of VideoRefer-700K Dataset and Benchmark

4.1. Human Evaluation on Reviewer

In our muliti-agent data engine, we introduce the Reviewer to address potential errors and mismatches, thereby ensuring the quality of our VideoRefer-700K dataset. To assess the effectiveness of the Reviewer, we conducted a manual evaluation of its outputs. We define the evaluation metrics as follows:

- TP (True Positives): Items that the Reviewer identified as relevant and accurate, which are confirmed to be true upon manual inspection.
- TN (True Negatives): Items that the Reviewer discarded

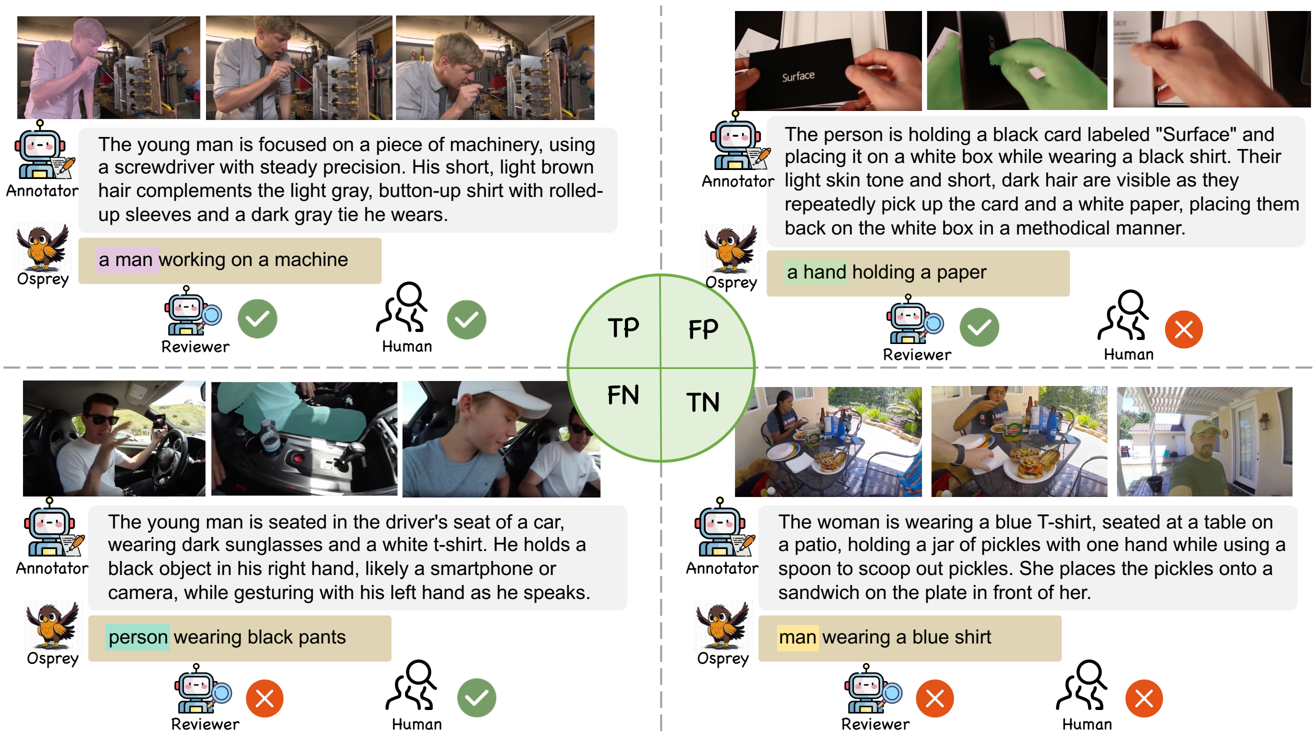


Figure A3. Visual illustrations of human check process. TP, TN, FP and FN are introduced for the assessment on Reviewer.

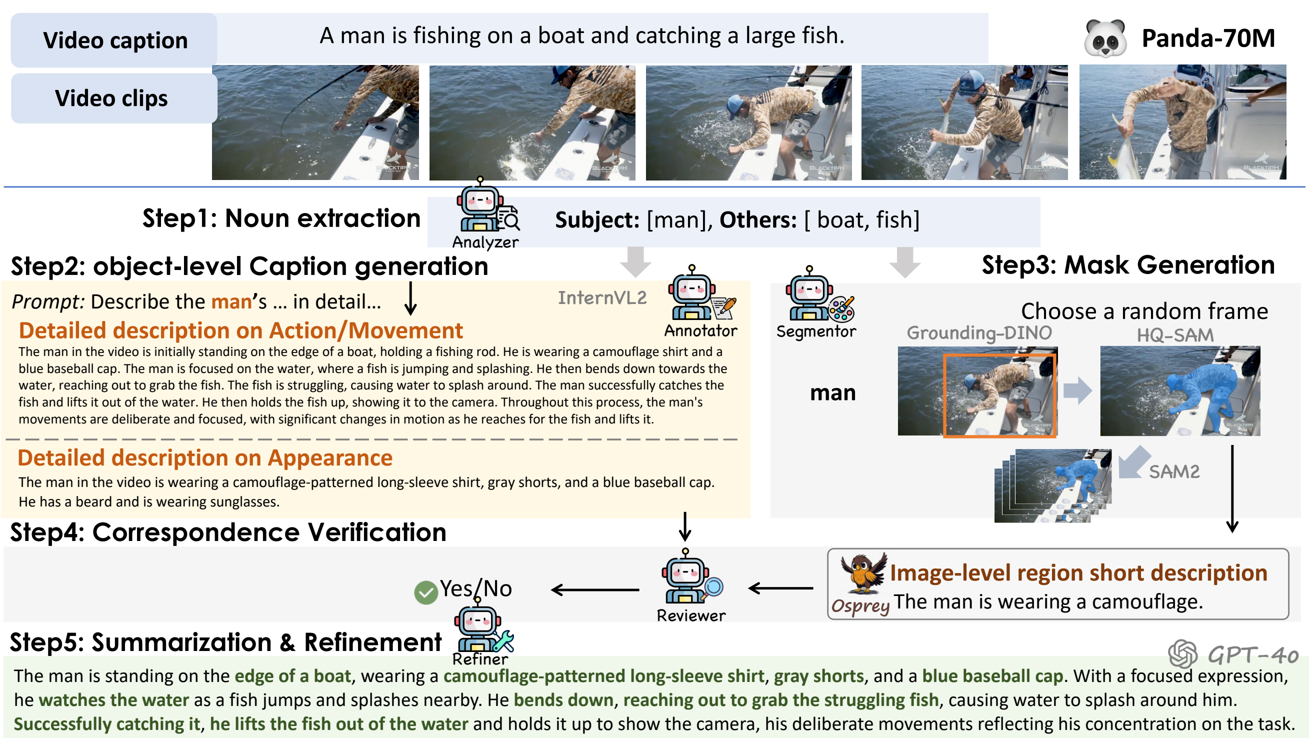


Figure A4. A detailed illustrative example of the construction pipeline in our multi-agent data engine.

	Manually True	Manually False
Reviewer True	88 (TP)	12 (FP)
Reviewer False	36 (FN)	64 (TN)

Table A5. Confusion matrix of the randomly sampled 100 items in the Reviewer evaluation.

as irrelevant or inaccurate, which are indeed false according to the manual check.

- FP (False Positives): Items that the Reviewer considered as true, but are found to be false during manual verification.
- FN (False Negatives): Items that the Reviewer discarded as false, but are actually true upon manual review.

We randomly sampled 100 items each from both the data discarded and retained by the Reviewer. The detailed results are represented in Table A5, and the corresponding metrics are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 0.76, \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} = 0.88, \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.71, \tag{3}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 0.79. \tag{4}$$

The precision value stands at 88%, indicating that the majority of samples identified as positive by the reviewer are indeed positive, thereby ensuring the data's quality.

4.2. Example Illustrations

We provide a typical example to better exhibit the construction pipeline of our multi-agent data engine, as shown in Fig. A4. Additionally, the data distribution of our VideoRefer-700K dataset is illustrated in Fig. A5. Fig. A7 further showcases the additional visual samples from the VideoRefer-700K dataset.

4.3. More Benchmark Visualization

We present more visualizations of our benchmark, VideoRefer-Bench$^{\mathrm{D}}$ and VideoRefer-Bench$^{\mathrm{Q}}$, as shown in Fig. A8. These visualizations aim to provide a deeper understanding of benchmarks' structure and content. VideoRefer-Bench$^{\mathrm{D}}$ focuses on detailed description tasks, facilitating the analysis of nuanced object references and relationships within videos. Meanwhile, VideoRefer-Bench$^{\mathrm{Q}}$ is designed for question-and-answer scenarios, capturing the essence of interactive video comprehension.

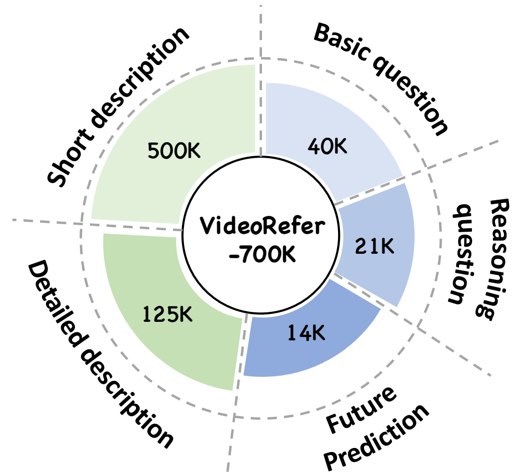


Figure A5. Data distributions of our VideoRefer-700K dataset, encompassing five different data types.

5. Limitations

In this work, our VideoRefer is designed on object-level spatial-temporal video understanding, without the abilities on grounding. This limitation may affect the applicability of our method in real-world scenarios, which requires identifying and associating objects within their dynamic contexts. In the future work, we will address this gap by integrating grounding abilities into our framework, extending our dataset and benchmark to improve the system's overall utility in practical applications.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv:2308.12966*, 2023. 1

[2] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *CVPR*, pages 12914–12923, 2024. 1

[3] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023.

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 1

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang

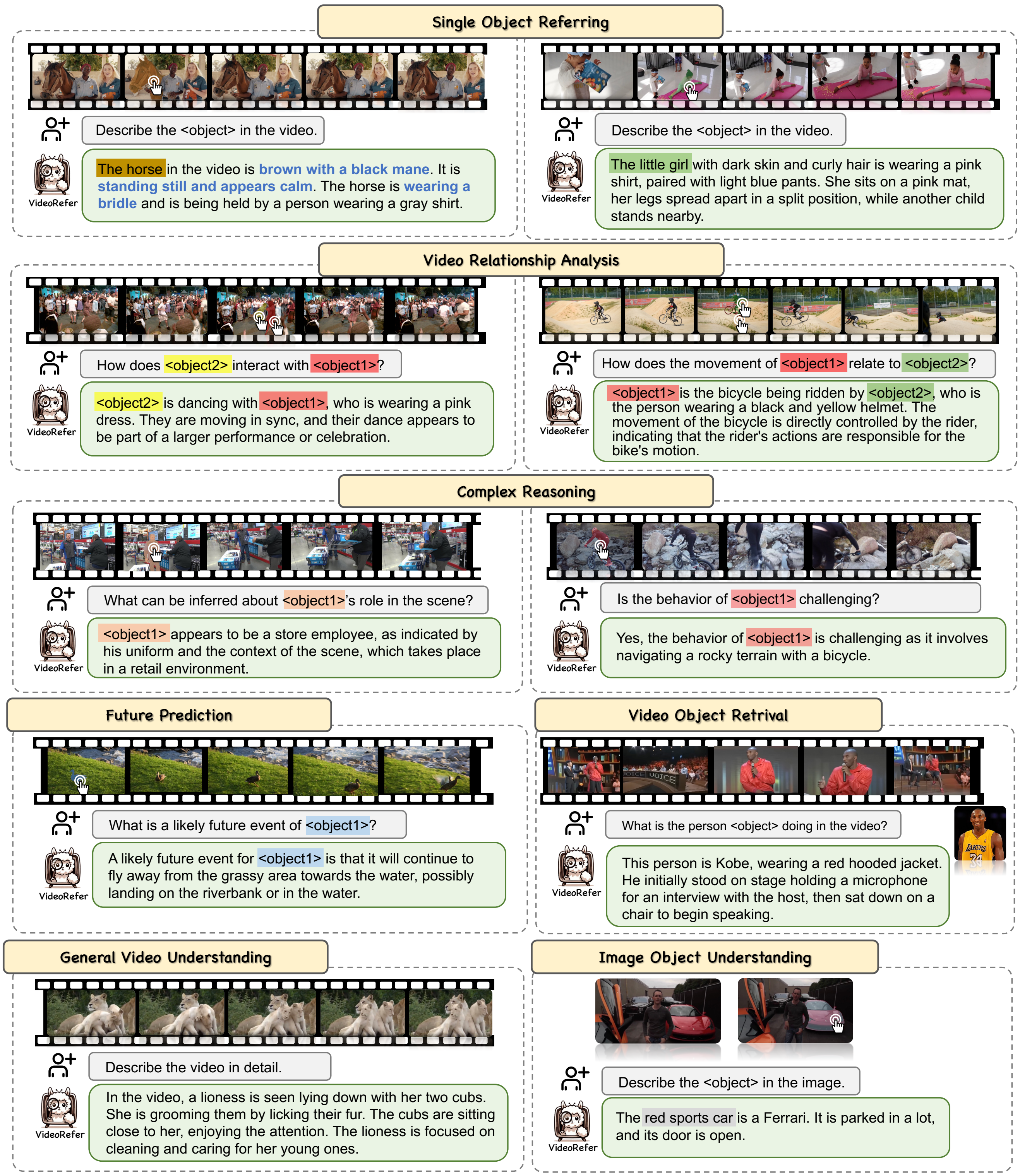
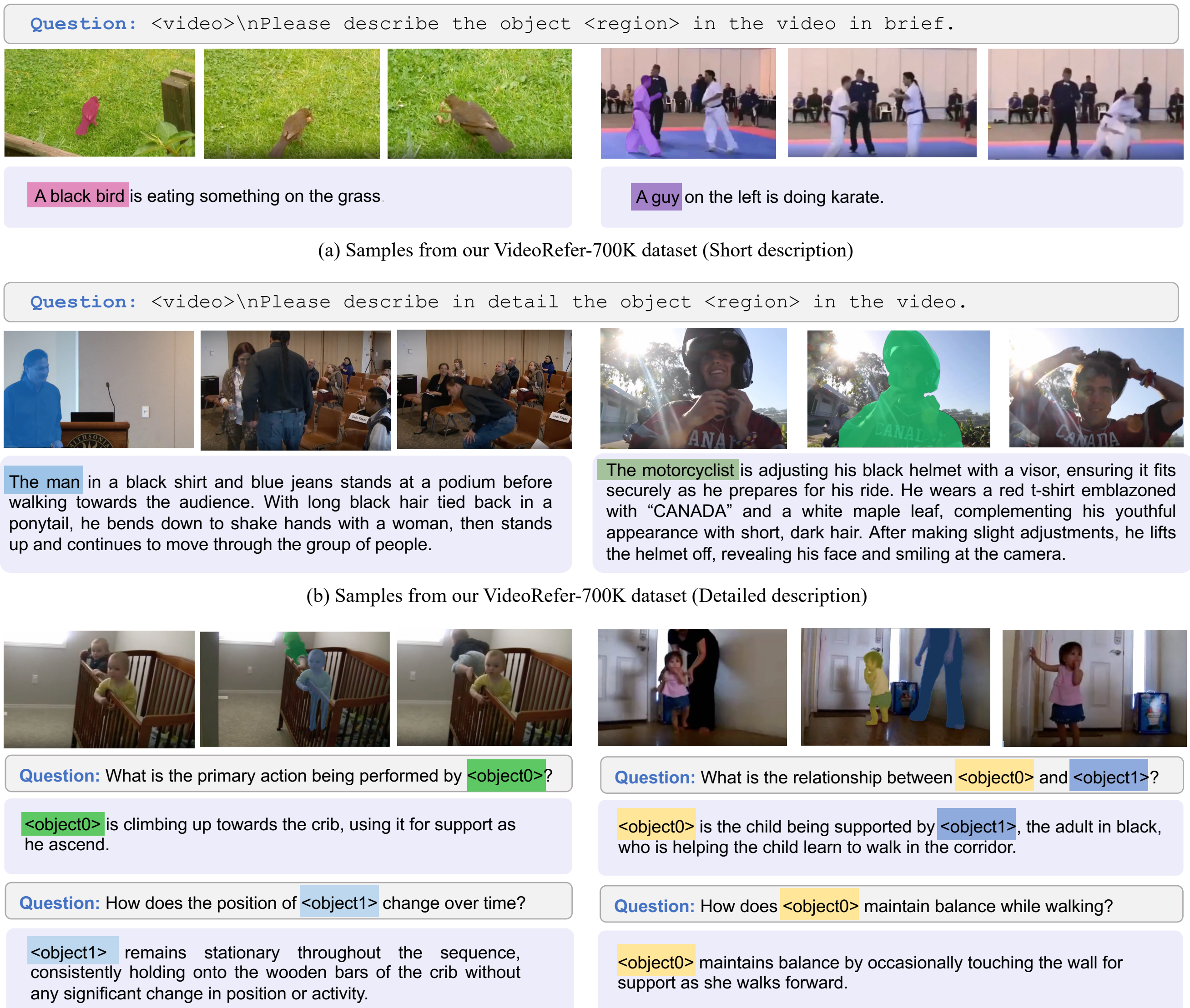


Figure A6. Visualization results of VideoRefer across various tasks, including single-object referring, video relationship analysis, complex reasoning, future prediction, video object retrieval, as well as general video understanding and image object understanding.



(a) Samples from our VideoRefer-700K dataset (Short description)

(b) Samples from our VideoRefer-700K dataset (Detailed description)

(c) Samples from our VideoRefer-700K dataset (QA)

Figure A7. Visual samples from our VideoRefer-700 dataset, typical including short descriptions, detailed descriptions, and QA pairs.

Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 3

[7] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *NeurIPS*, 2024. 1

[8] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *CVPR*, pages 13796–13806, 2024. 1

[9] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[10] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024. 1

[11] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024. 1

[12] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2023. 1

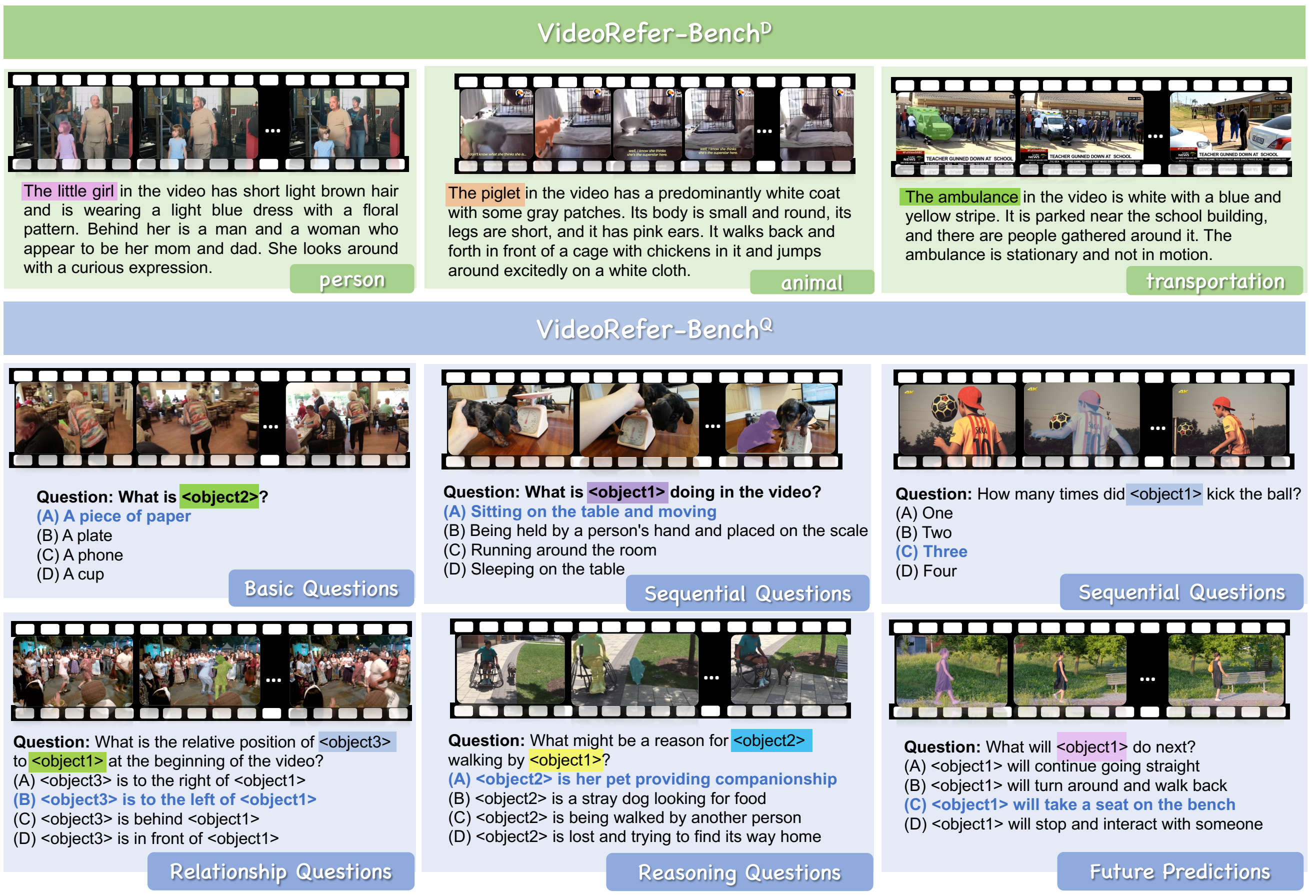


Figure A8. Visual examples of our VideoRefer-Bench, including VideoRefer-Bench^{D} and VideoRefer-Bench^{Q}.

[13] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 1

[14] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024. 1

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 3

[17] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024. 1

[18] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. 1

[19] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 1

[20] Jihao Qiu, Yuan Zhang, Xi Tang, Lingxi Xie, Tianren Ma, Pengyu Yan, David Doermann, Qixiang Ye, and Yunjie Tian. Artemis: Towards referential understanding in complex videos. In *NeurIPS*, 2024. 1

[21] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, pages 13009–13018, 2024. 1

[22] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 1

[23] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023. 1

[24] Yunjie Tian, Tianren Ma, Lingxi Xie, Jihao Qiu, Xi Tang, Yuan Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Chatterbox: Multi-round multimodal referring and grounding. *arXiv preprint arXiv:2401.13307*, 2024. 1

[25] Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *ECCV*, pages 166–185. Springer, 2024. 1

[26] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *CVPR*, pages 13838–13848, 2024. 1

[27] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024. 1

[28] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. In *ECCV*, pages 425–443. Springer, 2025. 1

[29] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *CVPR*, pages 28202–28211, 2024. 1

[30] Tongtian Yue, Jie Cheng, Longteng Guo, Xingyuan Dai, Zijia Zhao, Xingjian He, Gang Xiong, Yisheng Lv, and Jing Liu. Sc-tune: Unleashing self-consistent referential comprehension in large vision language models. In *CVPR*, pages 13073–13083, 2024.

[31] Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. *arXiv preprint arXiv:2403.09333*, 2024. 1

[32] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1

[33] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 1

[34] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 1

[35] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 1

[36] Wenqiao Zhang, Tianwei Lin, Jiang Liu, Fangxun Shu, Haoyuan Li, Lei Zhang, He Wanggui, Hao Zhou, Zheqi Lv, Hao Jiang, et al. Hyperllava: Dynamic visual and language expert tuning for multimodal large language models. *arXiv preprint arXiv:2403.13447*, 2024. 1

[37] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`, 2024. 1

[38] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474*, 2023. 1