Appendix

A Visualizations of Annotated Frames

Examples of annotated videos in ReVOS are shown in Fig. A.

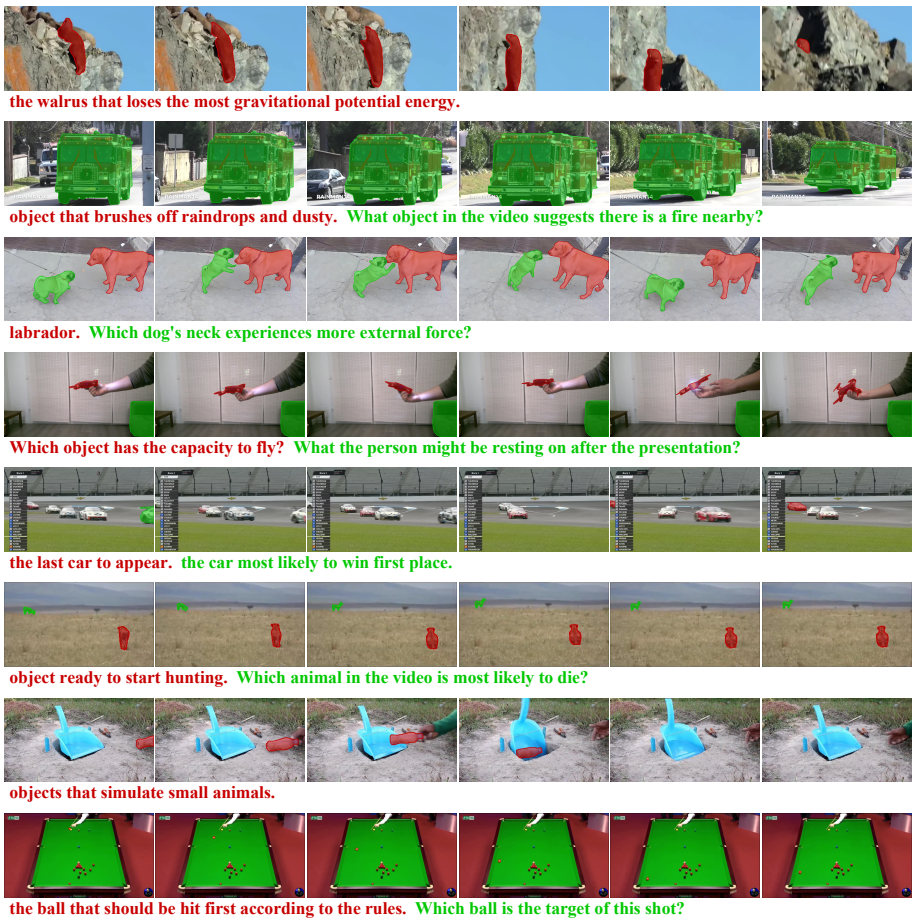

Fig. A: Sample videos in ReVOS.

B Implementation Details

We adopt the decoder in SAM [4] as the segmentation decoder. We choose Chat-UniVi-7B and Chat-UniVi-13B [3] as the pre-trained Multi-Modal LLM. The

number of visual tokens L per frame is set to 112, the same as the number in Chat-UniVi. We utilize XMem [1], a semi-supervised Video Object Segmentation method as the Object Tracker. The Text-guided Frame Sampler, Visual Backbone, and Object Tracker are all frozen during the training. Only the multi-modal LLM and SAM decoders are trainable. We leverage LoRA [2] to perform efficient fine-tuning of the multi-modal LLM. During training, we randomly sample a target frame f_{tgt} and 8-12 reference frames $\mathbf{x}_t$ per video instead of using the Text-guided Frame Sampler, to achieve more comprehensive training. During inference, we use the Text-guided Frame Sampler to obtain f_{tgt} and 12 reference frames $\mathbf{x}_t$ with the Global-Local sampling strategy. We use 8 NVIDIA 80G A100 GPUs for training. The training scripts are based on the deepspeed [6] engine. We train VISA for 10 epochs with a batch size of 128. Specifically, the batch size per device is set to 1, and the gradient accumulation step is set to 16, leading to 128 samples on 8 GPUs in total. We employ the AdamW [5] optimizer with a cosine schedule. The learning rate is set to 2e-5. All input frames are resized to 224×224 before feeding into the LLM, while the frame for the segmentation branch is resized to 1024×1024. The weights of the text generation loss λ_{txt} and the mask loss λ_{mask} are set to 1.0 and 1.0, respectively. The weights of the binary cross-entropy loss λ_{bce} and the dice loss λ_{dice} are set to 2.0 and 0.5, respectively. Check `https://github.com/cilinyan/VISA` for more implementation details.

References

1. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European Conference on Computer Vision. pp. 640–658. Springer (2022)
2. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
3. Jin, P., Takanobu, R., Zhang, C., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. arXiv preprint arXiv:2311.08046 (2023)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
5. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
6. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3505–3506 (2020)