

ViCaS: A Dataset for Combining Holistic and Pixel-level Video Understanding using Captions with Grounded Segmentation

Supplementary Material

A. Implementation Details

We provide the full set of implementation details and training hyperparameters for our Video-LLaVA-Seg model here.

Vision Backbone. Our vision backbone is a pretrained AM-RADIO [10] ViT-H/16 model. The video frames are resized by scaling the longer dimension to 384 in aspect-ratio preserving manner, followed by zero padding to obtain a square image with size 384×384 . The vision backbone has a stride of 16, thus yielding $24 \times 24 = 576$ tokens per frame. We sample a total of $T = 32$ frames per video. Of these $T_s = 8$ are encoded as slow frames. Meanwhile, all $T = 32$ fast frames are resized to $H_f = W_f = 4$ using adaptive average pooling, following the original work by Xu *et al.* [15]. Thus, each ‘fast’ frame is encoded using $4 \times 4 = 16$ tokens. Overall, the input video is represented as $N_v = (32 \times 16) + (8 \times 576) = 5120$ tokens at the input to the LLM.

Segmentation Network. The segmentation backbone is a Hiera-Small [12] with a Feature Pyramid Network (FPN) [9]. Since finegrained details are needed to predice accurate segmentation masks, we use a larger input resolution of 1024×1024 for the segmentation network, following the original implementation from Ravi *et al.* [11]. The backbone Hiera model has a stride of 16, thus resulting in feature maps of size 64×64 . The FPN yields two high-resolution feature maps at $8\times$ and $4\times$ strides, *i.e.* 128×128 and 256×256 , respectively. These feature maps are used in the final stages of the mask decoder [14] to predict high-resolution segmentation masks. The entire segmentation network (backbone, FPN, and mask decoder) is initialized with pretrained weights from SAM2 [11].

Training. Our Video-LLaVA-Seg model is trained in three stages:

- Stage 1: Pretraining stage where only the projection MLP is optimized for video captioning in order to align vision and language features.
- Stage 2: The projection MLP, vision backbone and LLM are optimized for video captioning.
- Stage 3: The entire model (projection MLP, vision backbone, LLM, and segmentation network) are optimized

Details about each stage of training are given in Table 1. Note that we only use a small fraction of the data from WebVid10M [3] and Panda70M [4]: for stage 1 we utilize 750,000 samples from each dataset, and for stage 2 we utilize 1,000,000 samples from each.

B. Grounded Captions to LG-VIS Prompts

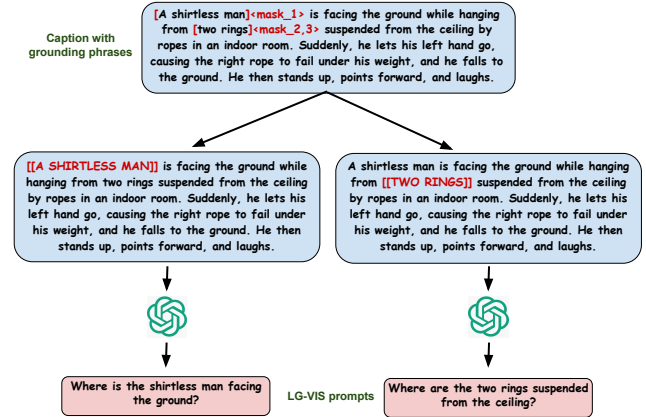


Figure 1. **Grounded captions to LG-VIS prompts.** We use GPT4 [1] to convert our human-written captions with phrase grounding to phrase-specific prompts for LG-VIS. Note that the text highlighted in red is for ease of visualization only.

As mentioned in Sec. 3.3 of the main text, our benchmark comprises a Language-Guided Video Instance Segmentation (LG-VIS) task which requires segmenting multiple objects based on a language prompt. To obtain these prompts from our grounded captions, we use a simple pipeline which is illustrated in Fig. 1. We input each grounding phrase to GPT4 with the phrase itself being highlighted using a special syntax as shown in the second row of the figure. GPT4 is given instructions to generate a ‘Where is?’ style question that references the object(s) in the grounding phrase. It is told to avoid putting too much information in the prompt and only include just enough information for the prompt to be unambiguous. The output is an LG-VIS prompt for each grounding phrase.

C. Dataset Statistics

The statistics for the train, validation and test sets of our ViCaS dataset are given in Table 2. Note that our test set has slightly higher object density than the train and validation sets, presenting a more challenging evaluation scenario.

D. Benchmark Results (Test Set)

We provide benchmark results for the test set in Table 3. We see that Video-LLaVA-Seg outperforms other baselines and existing task-specific approaches, with the LLaMA3-8B [6]

Stage	1	2	3
Tasks	VC	VC	VC, LG-VIS
Datasets	WebVid10M [3], Panda70M [4]	WebVid10M [3], Panda70M [4]	ViCaS, MeViS [5], Ref-YTVOS [13]
Epochs	1	1	10
Iterations	5,860	15,625	10,752
Batch Size	256	128	128
Optimized Components	Projection MLP	Projection MLP + Vision Backbone + LLM	Projection MLP + Vision Backbone + LLM + Segmentation Network
Learning Rate	1e-3	Vision backbone: 1e-6, Rest: 2e-5	Vision backbone: 1e-6, Rest: 2e-5

Table 1. **Training details for various stages.** VC: Video Captioning. LG-VIS: Language-Guided Video Instance Segmentations.

Split	Videos	Avg Duration (seconds)	Avg Caption (words)	Object Tracks	LG-VIS Prompts	Object Masks (Human)	Object Masks (Automatic)
Train	14,516	9.0	38.8	46,235	42,024	445,368	12.3M
Validation	2,950	8.7	38.2	9,265	8,393	87,054	2.4M
Test	2,950	9.8	40.6	10,088	9,019	104,661	2.9M
All	20,416	9.1	39.0	65,588	59,436	637,083	17.7M

Table 2. **Dataset statistics for train, validation and test splits.** As mentioned in Sec. 3.2, professional human annotators draw segmentation masks at 1fps, followed by using an off-the-shelf SAM2 [11] model to increase the temporal density to 30 fps. Both types of mask annotations are provided separately in the last two columns.

Model	CA	mAP	AP ₅₀	AP ₇₅	AP ₉₀
LLaVA-OV [8] (ZS)	2.9	-	-	-	-
MiniCPM-o 2.6 (ZS)	3.0	-	-	-	-
LMPM [5]	-	6.3	13.6	5.4	1.1
DsHmp [7]	-	10.4	22.3	8.9	2.0
VideoLISA [2]	-	7.8	17.3	6.2	1.3
Video-LLaVA-Seg	3.0	16.5	32.1	15.0	3.8

Table 3. Benchmark Results on our validation set. Refer to supplementary for test set results. CA: Caption Accuracy

backbone providing the highest performance, which is consistent with the trends seen on validation set. Compared to the validation set, the test set is more challenging from a segmentation perspective, evident from the lower scores for all methods on the LG-VIS task.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *Arxiv*, 2023. 1
- [2] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *NeurIPS*, 2024. 2
- [3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1, 2
- [4] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024. 1, 2
- [5] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 2
- [6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *Arxiv*, 2024. 1
- [7] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, 2024. 2
- [8] 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*, 2024. 2
- [9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1
- [10] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, 2024. 1
- [11] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *Arxiv*, 2024. 1, 2
- [12] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, 2023. 1
- [13] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos:

Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020. [2](#)

- [14] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. [1](#)
- [15] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *Arxiv*, 2024. [1](#)