General and Task-Oriented Video Segmentation Supplemental Material

The appendix is **structured** as follows:

- §A provides more implementation details of GvSeg.
- §B provides more experimental settings of GvSeg.
- §C shows additional quantitative results on YouTube-VIS$_{19}$ [1].
- §D broadly discusses the Limitation, Boarder Impact and Future Work.
- §E supplements more visualization results.

A More Implementation Details

Training. During training, for YouTube-VIS [1]/VOS [2], the input frames are randomly cropped to ensure that the longer side is at most 768p/1024p for ResNet/Swin backbones, respectively. The shorter side is resized to at least 240p/360p and at most 480p/600p for ResNet/Swin. For OVIS [3]/VSPW [4]/ VIPSeg [5]/KITTI [6]/ BURST [7], we resize the input frame so that the shorter side is at least 480p and at most 800p and the longer side is at most 1333p. The learning rate is scheduled following a step policy, decayed by a factor of 10 at 7K/11K for 10K/15K total training steps, respectively. Following existing solutions [8–11], we generate pseudo videos from MS COCO [12] as training samples for YouTube-VOS$_{18}$/YouTube-VIS$_{21}$ while no additional data is used for other benchmarks. We use standard data augmentations, *i.e.*, flipping, random scaling and cropping.
Testing. The evaluation process follows existing work [13–16] and adopts no test-time augmentation to ensure a fair comparison. For YouTube-VOS$_{18}$/YouTube-VIS$_{21}$, videos are resized to 360p/480p for ResNet/Swin backbones. For OVIS/VSPW/VIPSeg/KITTI/BURST, videos are tested at a resolution of 720p.
Reproducibility. GvSeg is implemented in PyTorch and trained on eight Tesla A40 GPUs. The testing is conducted on one Tesla A40 GPU.

B More Experimental Settings

Evaluation Metric for VPS. Following conventions [5, 6, 17], we adopt VPQ and STQ as metrics. VPQ computes the average panoptic quality from tube IoU across a span of several frames. For VIPSeg [5], we further report the VPQ scores for *thing* and *stuff* classes (*i.e.*, VPQ$^{\text{Th}}$ and VPQ$^{\text{St}}$). For KITTI-VPS [6], we

divide STQ into segmentation quality (SQ) and association quality (AQ) which evaluate the pixel-level tracking and segmentation performance in a video clip.
Evaluation Metric for VSS. Following the standard evaluation protocol [4, 13], we adopt the mean Intersection-over-Union (mIoU), and mean video consistency (mVC) which evaluates the category consistency among a video clip containing 8/16 frames (*i.e.*, mVC_8 and mVC_{16}) as metrics.
Evaluation Metric for VIS. Following the official setup [1, 3], we report the mean average precision (mAP) by averaging multiple IoU scores with thresholds from 0.5 to 0.95 at step 0.05, and the average recall (AR) given 1/10 segmented instances per video (*i.e.*, AR_1, AR_{10}). AP_{50} and AP_{75} with IoU thresholds at 0.5 and 0.75 are also employed for further analysis.
Evaluation Metric for EVS. For YouTube-VOS_{18}, we report region similarity ($\mathcal{J}$) and contour accuracy ($\mathcal{F}$) at *seen* and *unseen* classes. For BURST, we assess higher order tracking accuracy [18] on common (H_{com}) and uncommon (H_{unc}) classes.

C Additional Quantitative Results for VIS

We provide additional results on YouTube-VIS_{19} [1] in Table S1. YouTube-VIS_{19} consists of 2,238/343 videos for `train`/`val`. Following official setting [1, 3], we adopt mean average precision (mAP) and average recall (AR) as evaluation metrics. The training settings remain consistent with those used for YouTube-VIS_{21}. We observed that GvSeg consistently outperforms previous state-of-the-art methods in terms of mAP and AR.

D Discussion

Limitations. Although GvSeg has exhibited remarkable performance, environments with heavy occlusion and camera motion will result in subpar segmentation and tracking results. We show several representative failure cases in Fig. S8. We aim to address these limitations in our future work.
Broader Impact. Understanding visual scenes is a primary goal of computer vision. On the positive side, GvSeg represents a general video segmentation framework for EVS, VIS, VSS, and VPS which provides insight towards designing a universal model capable of addressing a broader spectrum of vision-related tasks. The disentanglement of task-specific properties of moving objects can benefit the wide application scenarios in video tasks such as Video Object Detection (VOD) and Multi-Object Tracking and Segmentation (MOTS). On the negative side, it's essential to acknowledge potential operational challenges our method may face in real-world applications. As a proactive step to mitigate any adverse effects on individuals and society, we advise the establishment of a robust security protocol which help ensure the safety and well-being of users and the broader community in case of any unforeseen issues.
Future Work. Following the basic idea to disentangle task-specific properties of instances in a dynamic video, we will extend GvSeg towards a universal model

Table S1: Quantitative results on YouTube-VIS$_{19}$ [1] `val` (§C).

Method	Backbone	Gen. Sol	mAP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$
MaskTrack [1]	R-50	✗	30.3	51.1	32.6	31.0	35.5
SipMask [19]	R-50	✗	33.7	54.1	35.8	35.4	40.1
CrossVIS [20]	R-50	✗	36.3	56.8	38.9	35.6	40.7
InsPro [11]	R-50	✗	37.6	58.7	0.9	32.7	41.4
VISOLO [21]	R-50	✗	38.6	56.3	43.7	35.7	42.5
InstMove [22]	R-50	✗	40.6	67.2	45.1	35.0	48.2
SeqFormer [9]	R-50	✗	47.4	69.8	51.8	45.4	54.8
MinVIS [23]	R-50	✗	47.4	69.0	52.1	45.7	55.7
IDOL [8]	R-50	✗	49.5	74.0	52.9	47.7	58.7
VITA [24]	R-50	✗	49.8	72.6	54.5	49.4	61.0
GenVIS [10]	R-50	✗	50.0	71.5	54.6	49.5	59.7
TCOVIS [25]	R-50	✗	49.5	71.2	53.8	41.3	55.9
CTVIS [26]	R-50	✗	50.1	73.7	54.7	41.8	59.5
Mask2Former [27]	R-50	✓	46.4	68.0	50.0	-	-
CAROQ [28]	R-50	✓	46.7	70.4	50.9	45.7	55.9
TubeFormer [17]	R-50	✓	47.5	68.7	52.1	50.2	59.0
Tube-Link [13]	R-50	✓	52.8	75.4	56.5	49.3	59.9
GvSeg	R-50	✓	**54.9**	**76.6**	**60.1**	**50.6**	**63.0**

with shared weights in our future work. We aim to cover more video instance perception tasks such as accommodate Single Object Tracking (SOT), Multi-Object Tracking and Segmentation (MOTS), Referring Expression Segmentation (RES), and Video Object Detection (VOD), all while maintaining shared weights across these tasks. This endeavor signifies a step towards a foundation model of video perception. In addition, while GvSeg emphasizes a unified architecture, the prospect of unified training is promising, and we shall consider it as our future direction.

E Further Qualitative Results

In this section, we provide more qualitative results on five datasets, including OVIS [3] in Fig. S1, YouTube-VIS$_{21}$ [1] in Fig. S2, VSPW [4] in Fig. S3, BURST [7] in Fig. S4, YouTube-VOS [2] in Fig. S5 VIPSeg [5] in Fig. S6, and KITTI [6] in Fig. S7. We observe that GvSeg is able to produce highly exquisite results compared with previous competitive methods TarVIS [14] and Tube-Link [13].

References

1. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019) 1, 2, 3, 7
2. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018) 1, 3, 9
3. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H., Bai, S.: Occluded video instance segmentation: A benchmark. IJCV **130**(8), 2022–2039 (2022) 1, 2, 3, 6, 11
4. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: Vspw: A large-scale dataset for video scene parsing in the wild. In: CVPR (2021) 1, 2, 3, 8
5. Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., Yang, Y.: Large-scale video panoptic segmentation in the wild: A benchmark. In: CVPR (2022) 1, 3, 10, 11
6. Weber, M., Xie, J., Collins, M., Zhu, Y., Voigtlaender, P., Adam, H., Green, B., Geiger, A., Leibe, B., Cremers, D., et al.: Step: Segmenting and tracking every pixel. In: NeurIPS (2021) 1, 3, 11
7. Athar, A., Luiten, J., Voigtlaender, P., Khurana, T., Dave, A., Leibe, B., Ramanan, D.: Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In: WACV (2023) 1, 3, 9
8. Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., Bai, X.: In defense of online models for video instance segmentation. In: ECCV (2022) 1, 3
9. Wu, J., Jiang, Y., Bai, S., Zhang, W., Bai, X.: Seqformer: Sequential transformer for video instance segmentation. In: ECCV (2022) 1, 3
10. Heo, M., Hwang, S., Hyun, J., Kim, H., Oh, S.W., Lee, J.Y., Kim, S.J.: A generalized framework for video instance segmentation. In: CVPR (2023) 1, 3
11. He, F., Zhang, H., Gao, N., Jia, J., Shan, Y., Zhao, X., Huang, K.: Inspro: Propagating instance query and proposal for online video instance segmentation. In: NeurIPS (2022) 1, 3
12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 1
13. Li, X., Zhang, W., Pang, J., Chen, K., Cheng, G., Tong, Y., Loy, C.C.: Tube-link: A flexible cross tube baseline for universal video segmentation. In: ICCV (2023) 1, 2, 3
14. Athar, A., Hermans, A., Luiten, J., Ramanan, D., Leibe, B.: Tarvis: A unified approach for target-based video segmentation. In: CVPR (2023) 1, 3
15. Yan, B., Jiang, Y., Wu, J., Wang, D., Yuan, Z., Luo, P., Lu, H.: Universal instance perception as object discovery and retrieval. In: CVPR (2023) 1
16. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In: NeurIPS (2021) 1
17. Kim, D., Xie, J., Wang, H., Qiao, S., Yu, Q., Kim, H.S., Adam, H., Kweon, I.S., Chen, L.C.: Tubeformer-deeplab: Video mask transformer. In: CVPR (2022) 1, 3
18. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. IJCV **129**, 548–578 (2021) 2
19. Cao, J., Anwer, R.M., Cholakkal, H., Khan, F.S., Pang, Y., Shao, L.: Sipmask: Spatial information preservation for fast image and video instance segmentation. In: ECCV (2020) 3
20. Yang, S., Fang, Y., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Crossover learning for fast online video instance segmentation. In: ICCV (2021) 3

21. Han, S.H., Hwang, S., Oh, S.W., Park, Y., Kim, H., Kim, M.J., Kim, S.J.: Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In: CVPR (2022) 3
22. Liu, Q., Wu, J., Jiang, Y., Bai, X., Yuille, A.L., Bai, S.: Instmove: Instance motion for object-centric video segmentation. In: CVPR (2023) 3
23. Huang, D.A., Yu, Z., Anandkumar, A.: Minvis: A minimal video instance segmentation framework without video-based training. In: NeurIPS (2022) 3
24. Heo, M., Hwang, S., Oh, S.W., Lee, J.Y., Kim, S.J.: Vita: Video instance segmentation via object token association. In: NeurIPS (2022) 3
25. Li, J., Yu, B., Rao, Y., Zhou, J., Lu, J.: Tcovis: Temporally consistent online video instance segmentation. In: ICCV (2023) 3
26. Ying, K., Zhong, Q., Mao, W., Wang, Z., Chen, H., Wu, L.Y., Liu, Y., Fan, C., Zhuge, Y., Shen, C.: Ctvis: Consistent training for online video instance segmentation. In: ICCV (2023) 3
27. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021) 3
28. Choudhuri, A., Chowdhary, G., Schwing, A.G.: Context-aware relative object queries to unify video instance and panoptic segmentation. In: CVPR (2023) 3

Fig. S1: More **visual comparison** for Video Instance Segmentation on OVIS [3].

Fig. S2: More **visual comparison** for Video Instance Segmentation on YouTube-VIS$_{21}$ [1].

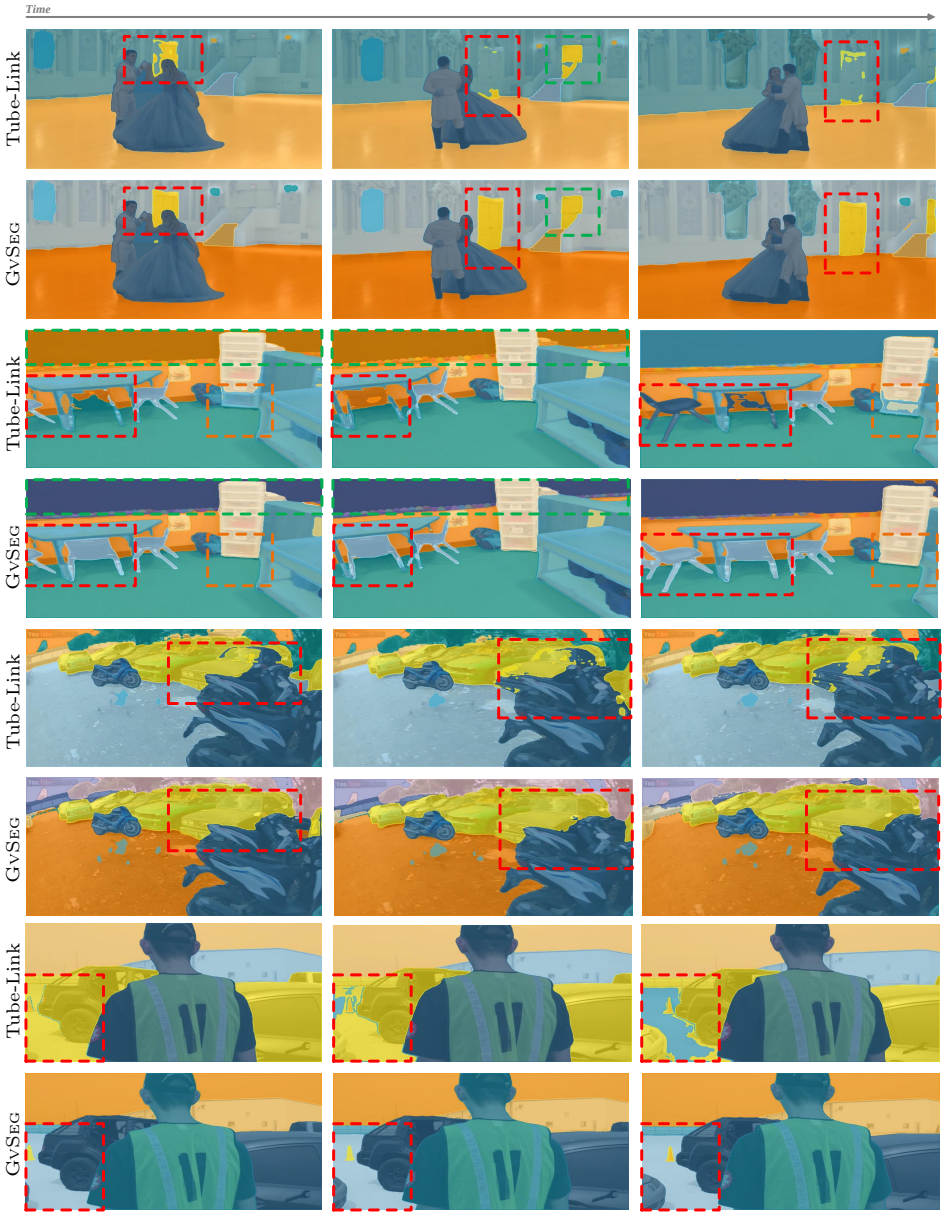

Fig. S3: More **visual comparison** for Video Semantic Segmentation on VSPW [4].

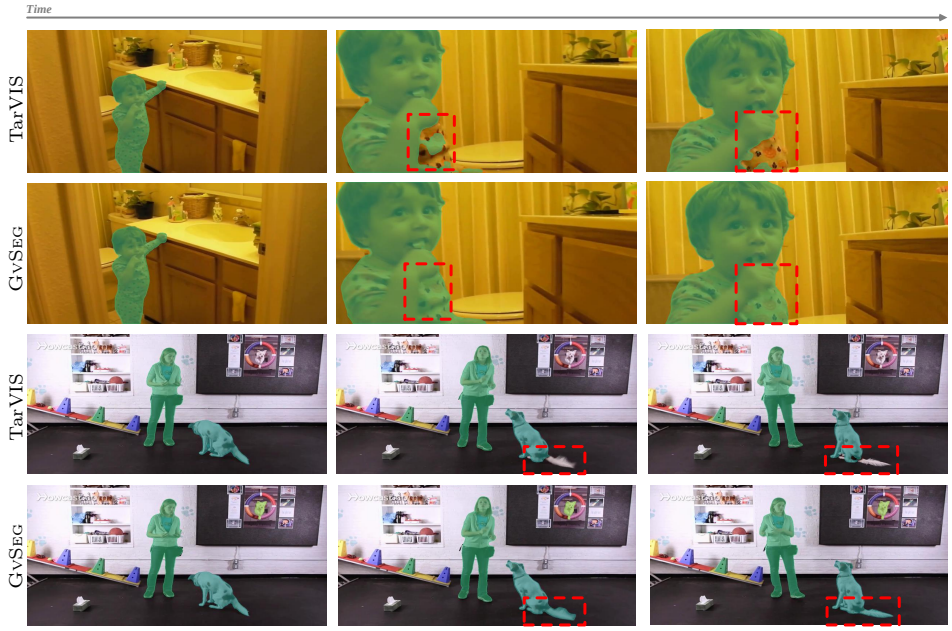

Fig. S4: Comparison for Exemplar-guided Video Segmentation on BURST [7].

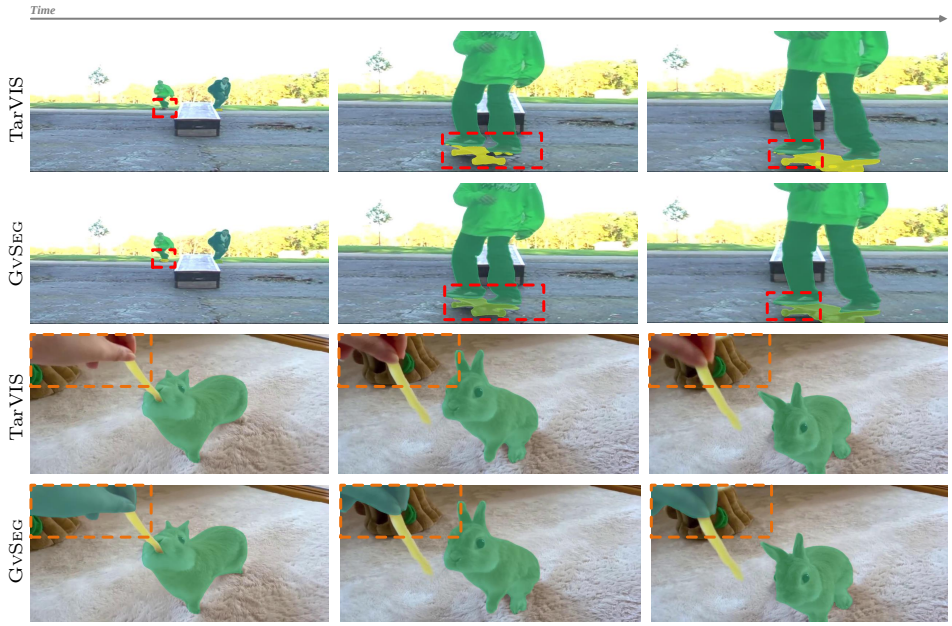

Fig. S5: Comparison for Exemplar-guided Video Segmentation on YouTube-VOS [2].

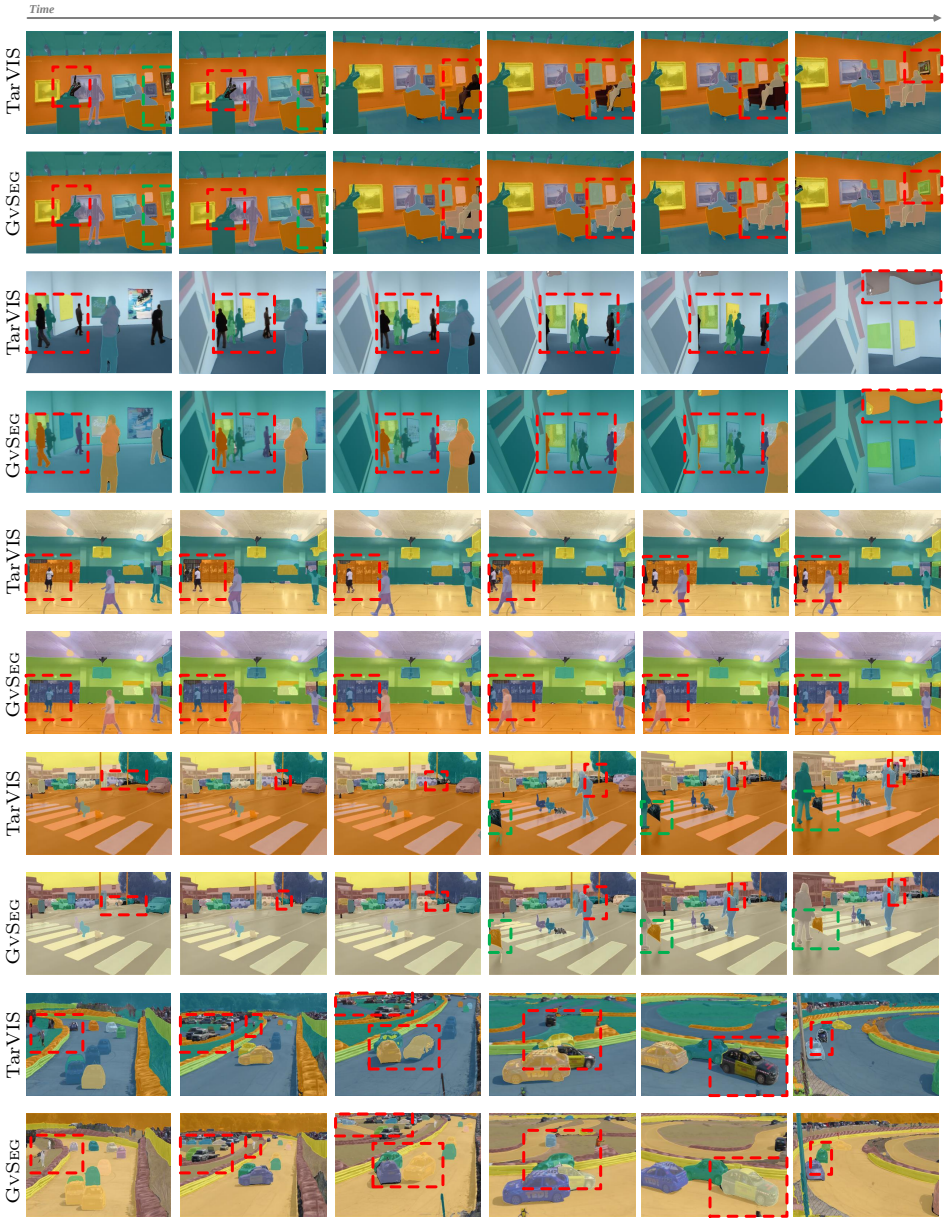

Fig. S6: More **visual comparison** for Video Panoptic Segmentation on VIPSeg [5].

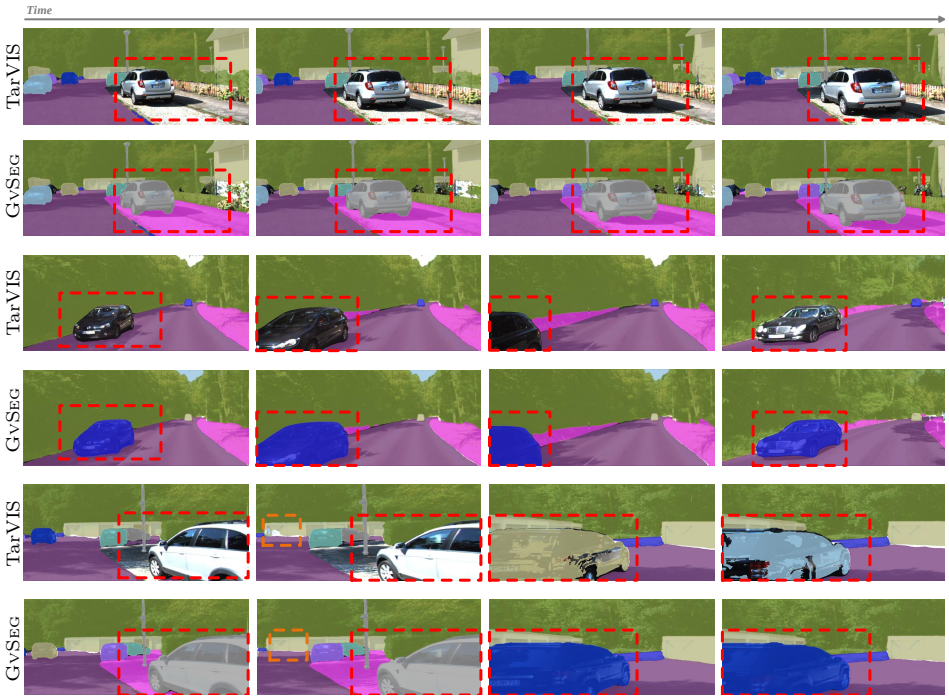

Fig. S7: More **visual comparison** for Video Panoptic Segmentation on KITTI [6].

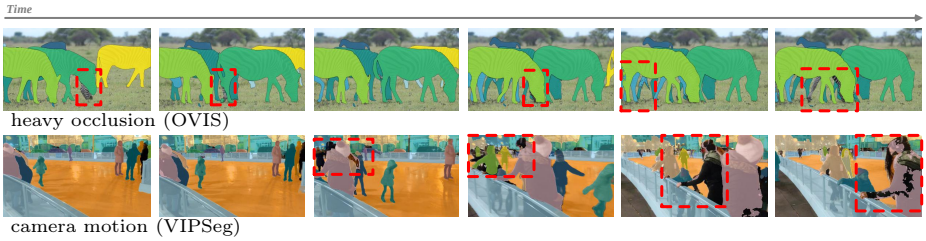

Fig. S8: Failure cases due to on OVIS [3] and VIPSeg [5]. See more details in §D.