Exploiting Temporal State Space Sharing for Video Semantic Segmentation

Supplementary Material

This supplementary material provides more results that enhance and extend the findings presented in the main manuscript. Due to space constraints, certain details and experiments were omitted from the primary manuscript. Specifically, Sec. A presents more ablation studies that offer deeper insights into the proposed TV3S model. Sec. B details the latest performance results substantiating the efficacy of our proposed method through a more fair and refined training procedure. Sec. C showcases an expanded set of visual results demonstrating the segmentation capabilities of TV3S, alongside comparative analyses with additional models including MRCFA.

A. Additional Ablation Studies

Following the main text, all ablation studies were conducted on the VSPW dataset using the MiT-B1 and Swin-T backbones, adhering to the same training and inference strategies outlined in the main text.

Effect of spatial information extraction. To assess the effectiveness of our proposed TV3S architecture in extracting spatial information, we conducted experiments using single-frame inputs and compared the performance against baseline segmentation models and other video semantic segmentation (VSS) methods, as presented in Tab. S1. While VSS methods are inherently designed for multi-frame processing, this evaluation isolates their ability to handle spatial features independently. For a fair comparison, we evaluated our model with and without the TV3S blocks, noting that our architecture can utilize the temporal blocks even when only one frame is provided. The results demonstrate that our model not only performs on par with the baseline when the TV3S blocks are excluded but also significantly outperforms it when the blocks are included. In contrast, other VSS methods exhibit reduced performance in single-frame evaluations, reflecting their ability to partially adapt to single-frame inputs despite their multi-frame design. These findings indicate that our TV3S model effectively captures spatial information and maintains robust performance even without temporal context, showcasing its superiority in both spatial and spatiotemporal segmentation tasks.

Effect of the number of TV3S blocks. As detailed in the main text, the MiT-B1 backbone exhibited enhanced performance with an increasing number of TV3S blocks, achieving a mIoU of 40.0 and improved temporal consistency metrics (mVC_8 = 90.7, mVC_{16} = 87.0) when utilizing four blocks, as shown in Tab. S3. Extending this evaluation to the Swin-T backbone and maintaining a consistent framework, the Swin-T backbone attained a mIoU of 44.90 with four blocks, closely aligning with its peak performance of 45.11 achieved using two blocks. Additionally, temporal consistency metrics (mVC_8 = 88.0, mVC_{16} = 83.5) remained stable across different block configurations. These findings indicate that, while the MiT-B1 backbone benefits significantly from an increased number of TV3S blocks, the Swin-T backbone maintains robust performance with a standardized four-block setup, underscoring the effectiveness of a unified framework for diverse backbones.

Methods	Backbones	mIoU↑	WIoU
Segformer	MiT-B1	36.5	58.8
CFFM	MiT-B1	37.1	59.0
MRCFA	MiT-B1	37.0	58.8
TV3S (Ours)	MiT-B1	37.7	59.2
TV3S (+Blocks)	MiT-B1	**38.6**	**60.3**
Segformer	MiT-B2	43.9	63.7
CFFM	MiT-B2	43.6	63.3
MRCFA	MiT-B2	43.4	63.5
TV3S (Ours)	MiT-B2	43.8	62.8
TV3S (+Blocks)	MiT-B2	**44.9**	**63.7**
Segformer	MiT-B5	48.9	65.1
CFFM	MiT-B5	48.3	65.8
MRCFA	MiT-B5	48.0	65.3
TV3S (Ours)	MiT-B5	48.9	66.0
TV3S (+Blocks)	MiT-B5	**49.5**	**66.4**
Mask2Former	Swin-T	41.2	62.6
TV3S (Ours)	Swin-T	42.8	62.4
TV3S (+Blocks)	Swin-T	**43.8**	**62.6**
Mask2Former	Swin-S	42.1	63.1
TV3S (Ours)	Swin-S	49.5	65.8
TV3S (+Blocks)	Swin-S	**50.5**	**66.2**

Table S1. Comparative effectiveness of models in extracting spatial information from single-frame inputs on the VSPW dataset, with our proposed method outperforming existing models.

Training with different temporal context. We assessed the impact of varying the number of template frames during training on the MiT-B1 backbone variant of TV3S, as detailed in Tab. S4. Specifically, the model was trained with one ($\{I_{t-3}, I_t\}$), two ($\{I_{t-6}, I_{t-3}, I_t\}$), three ($\{I_{t-9}, I_{t-6}, I_{t-3}, I_t\}$) and five ($\{I_{t-15}, I_{t-12}, I_{t-9}, I_{t-6}, I_{t-3}, I_t\}$) template frames. The results indicate a clear improvement in visual consistency as the number of templates increases, showcasing the model's enhanced ability to maintain temporal coherence, attributed to the specialized training methodology. However, while using five templates yielded the highest mVC values, the mIoU performance peaked with three templates, offering a balanced trade-off between segmentation accuracy and

Methods	Backbones	mIoU↑	mVC_8↑	mVC_{16}↑	GFLOPs↓	Params(M)↓	FPS↑
Mask2Former	R50	38.5	81.3	76.4	110.6	44.0	19.4
MPVSS	R50	37.5	84.1	77.2	38.9	84.1	33.9
Mask2Former	R101	39.3	82.5	77.6	141.3	63.0	16.9
MPVSS	R101	38.8	84.8	79.6	45.1	103.1	32.3
DeepLabv3+	R101	34.7	83.2	78.2	379.0	62.7	9.2
UperNet	R101	36.5	82.6	76.1	403.6	83.2	16.0
PSPNet	R101	36.5	84.2	79.6	401.8	70.5	13.8
OCRNet	R101	36.7	84.0	79.0	361.7	58.1	14.3
TCB	R101	37.8	87.9	84.0	1692	-	-
ETC	OCRNet	37.5	84.1	79.1	361.7	-	-
Segformer	MiT-B5	48.9	87.8	83.7	185.0	82.1	9.4
CFFM	MiT-B5	49.3	90.8	87.1	413.5	85.5	4.5
MRCFA	MiT-B5	49.9	90.9	87.4	373.0	84.5	5.0
TV3S (Ours)	MiT-B5	**50.4**	**91.9**	**89.1**	137.0	85.6	14.0

Table S2. **Updated** quantitative comparison of our MiT-B5 model with existing methods on the VSPW dataset. Our model achieves the best balance among *accuracy*, *model complexity*, and *operational speed*. FPS and FLOPs are calculated with an input resolution of 480 × 853.

Backbones	TV3S Blocks	mIoU	mVC_8	mVC_{16}
MiT-B1	1	38.4	88.3	83.7
	2	39.2	89.5	85.3
	3	39.6	88.7	84.2
	4	**40.0**	**90.7**	**87.0**
Swin-T	1	44.66	87.9	83.3
	2	**45.11**	88.4	83.9
	3	44.41	88.3	83.8
	4	**44.90**	88.0	83.5

Table S3. Performance metrics based on the number of TV3S blocks in the model.

Templates No.	mIoU	mVC_8	mVC_{16}
1	38.1	90.3	83.6
2	37.6	90.5	84.3
3	**40.0**	90.7	87.0
5	38.1	**91.2**	**88.0**

Table S4. Evaluation based on the number of templates exposed during training.

Models	Evaluation (mIoU)		
	Bi	Bi+Embed	Direct
1 TSS (No Shift)	37.33	38.0	38.0
TV3S (No Shift)	38.0	38.4	38.9
TV3S (Shift)	**39.6**	37.6	39.5

Table S5. Implications of using bi-directional representation with embedding on the proposed architecture.

Methods	Backbones	mIoU	mVC_8	mVC_{16}
VideoMamba	MiT-B1	36.2	83.9	78.7
TV3S	MiT-B1	40.0	90.7	87.0
MPVSS	Swin-B	52.6	89.5	85.9
MPVSS	Swin-L	53.9	89.6	85.8
TV3S	Swin-B	53.0	90.3	86.8
TV3S	Swin-L	55.6	90.7	87.5

Table S6. Additional Experiments with VideoMamba as decoder and with bigger Swin Transformer backbones

temporal consistency. Although further fine-tuning could refine the model for specific scenarios, the configuration with three templates is recommended for its optimal balance, aligning with findings from. This configuration ensures the model operates effectively within practical constraints while leveraging its temporal modeling strengths.

Applicability of Bi-directional Scanning. We investigated the use of bi-directional scanning, a technique prevalent in recent vision-based approaches utilizing mamba, in the MiT-B1 variant of TV3S (see Tab. S5). This method involved scanning the encoded feature space in both directions, with or without adding embeddings during the scanning process, effectively doubling the computational load for the decoder. The experimental results indicated that incorporating bi-directional scanning did not enhance performance and, in some cases, led to degradation. We believe that this decline may be due to two factors: first, the implementation was conducted in a pixel-wise manner within the encoded feature space, differing from the patch-wise approach in the original mamba implementations; second, scanning the same feature space twice might disrupt the continuity of information, potentially hindering the model's ability to maintain performance. Consequently, these findings suggest that while bi-directional scanning is effective in certain contexts, its application as a decoder in the present architecture did not yield benefits and may require further methodological refinements.

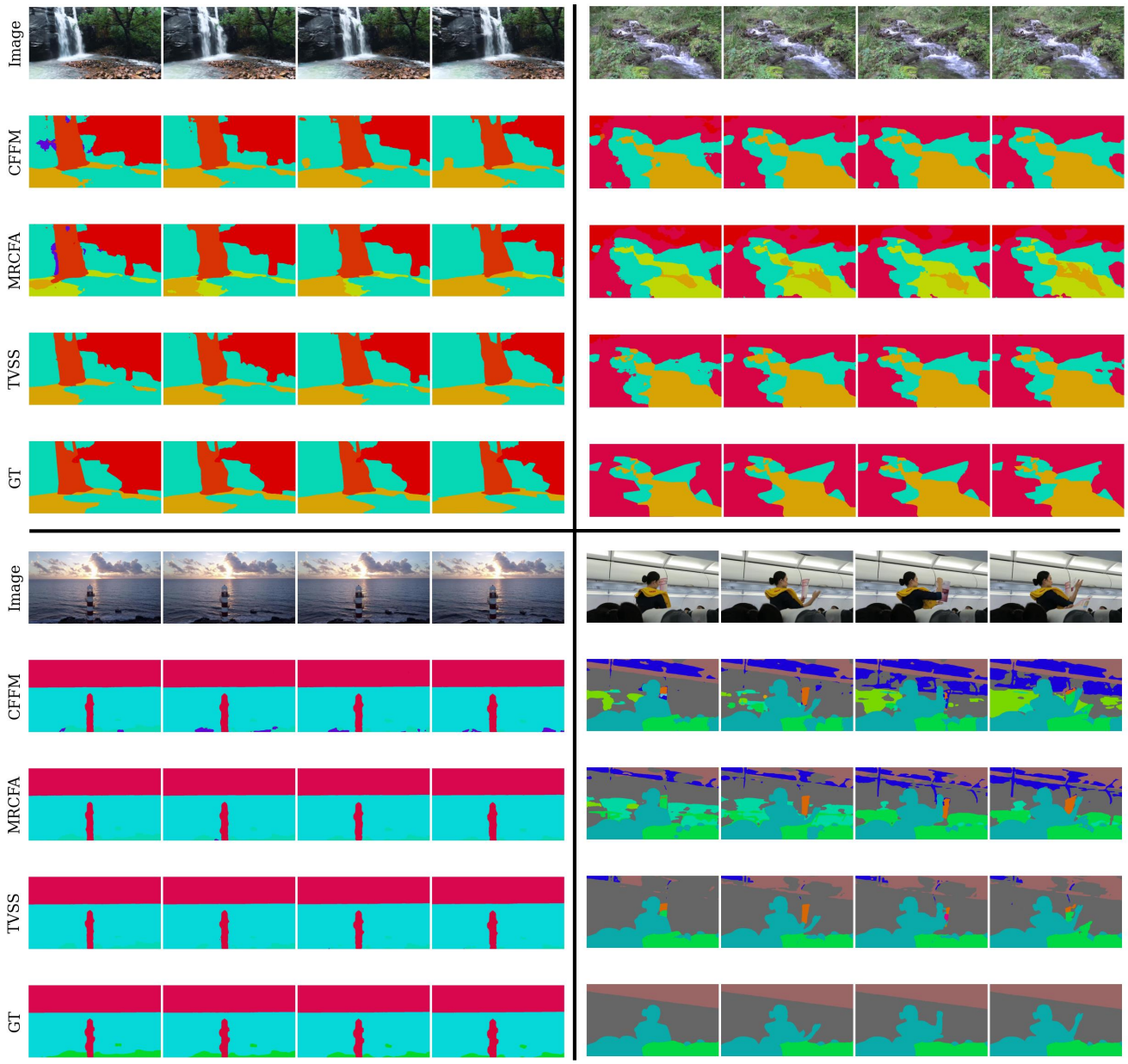

Figure S1. Additional examples showcasing the performance of the proposed TVSS architecture compared with other VSS methods, demonstrating visual consistency and accuracy.

Additional Experiments. Extended experiments were conducted during the rebuttal phase, which included testing VideoMamba and larger backbones of Swin, specifically its Swin-B and Swin-L variants, as tabulated in Tab. S6. For the experiments with VideoMamba, we used the MiT-B1 backbone in conjunction with VideoMamba as the decoder. It was observed that VideoMamba only achieved a mean Intersection over Union (mIoU) of 36.24, while our TV3S framework achieved an mIoU of 40.0, thanks to its effective state propagation and shifted-window mechanism, making it ideal for dense prediction tasks.

As for the experiments involving larger backbones, it was noted that by directly extending the current framework without hyper-parameter tuning, we achieved mIoU scores that are better than the performance of MPVSS. This finding highlights the robustness of our approach and ensures fair comparisons with other methods.

B. Updated Performance

Our initial training setup for the TV3S architecture, based on the `MMSegmentation` codebase, utilized two A100

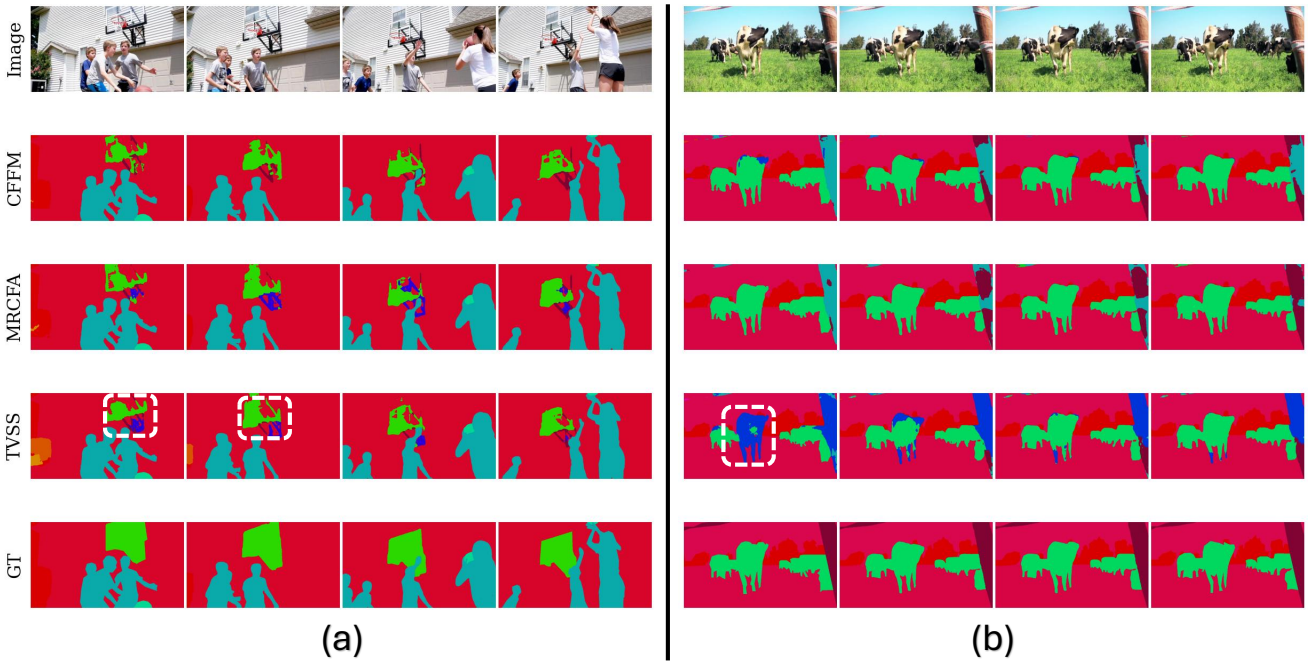

Figure S2. Failure cases of the proposed method: (a) errors in the presence of transparent objects and (b) initial segmentation errors propagating temporarily before being corrected.

NVIDIA GPUs with a batch size of 2 and trained the model for 160k iterations using three reference frames. This configuration resulted in strong temporal consistency metrics (mVC_8 and mVC_{16}), achieving a good trade-off between computational efficiency and frames per second (FPS). However, compared to other video semantic segmentation (VSS) methods that were trained using four GPUs, our model was exposed to fewer data variants, potentially impacting its generalization capabilities.

To ensure a fairer comparison, we extended the training duration by an additional 80k iterations, totalling 240k iterations—a 50% increase in training time. This adjustment compensates for the advantages other methods gain from using more GPUs, such as exposure to a wider variety of data and improved generalization. Concurrently, we halved the learning rate to 3e-5 from 6e-5 to maintain effective learning without overshooting, keeping the optimizer and learning rate scheduler configurations consistent.

Under this training setup, as shown in Tab. S2 , our proposed TV3S architecture achieved state-of-the-art performance across all evaluated metrics, including mean Intersection over Union (mIoU) and temporal consistency metrics (mVC_8 and mVC_{16}).

C. Additional Qualitative Examples

In this section, we present qualitative examples to further demonstrate the effectiveness of the proposed TVSS architecture. As shown in Fig. S1, the segmentation outputs from TVSS are compared with those from other state-of-the-art video semantic segmentation (VSS) methods. The examples illustrate how TVSS maintains good visual consistency across consecutive frames while achieving accurate segmentation. These results underline the advantages of the temporal state-sharing mechanism, which effectively propagates temporal information and reduces inconsistencies commonly observed in other methods. The visualizations in Fig. S1 provide a clear, comparative insight into how TVSS handles challenging scenarios, reinforcing the quantitative results discussed earlier.

C.1. Success Cases

The proposed TVSS architecture excels in ensuring both stability and continuity in the segmentation process across frames, maintaining a high level of consistency even in dynamic and complex environments. The following examples demonstrate the architecture's ability to preserve these qualities in challenging visual sequences.

(a) Temporal continuity and object consistency: One of the standout features of TVSS is its ability to maintain temporal continuity. In the provided sequences, the model shows a consistent and stable segmentation of dynamic objects, such as waterfalls, people, or animals, across multiple frames. This is particularly evident in cases where objects remain in motion or where the background changes slightly, but the segmentation boundaries remain stable, offering a

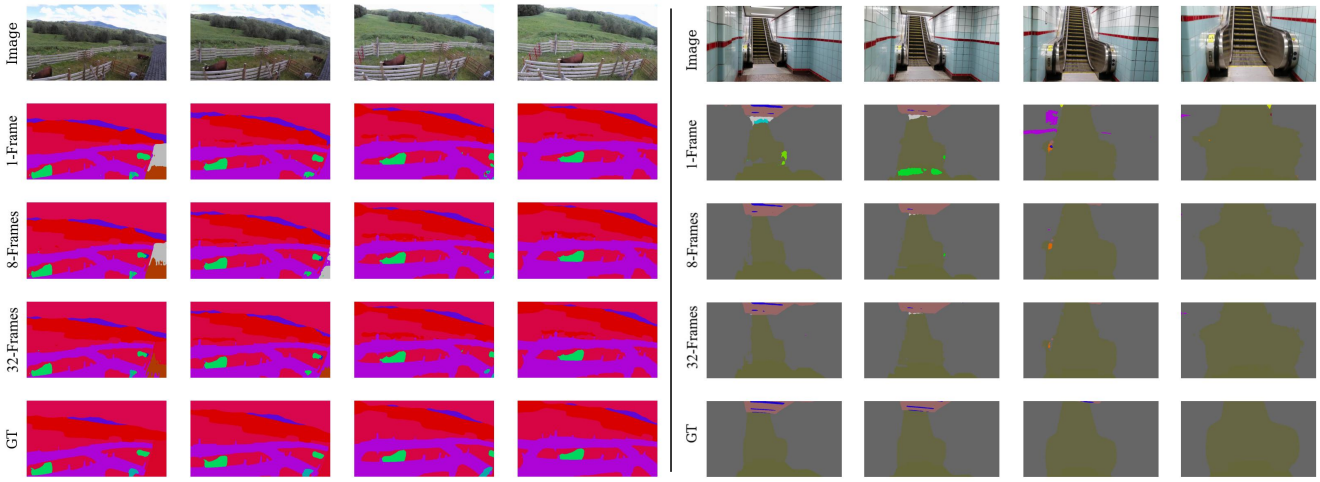

Figure S3. Visual comparison of segmentation results with 1, 8, and 32 exposed frames during inference.

smooth transition between frames.

(b) Robust segmentation in variable environments: In more challenging scenes, including those with changing lighting or background complexity, TVSS continues to show visual stability. The segmentation boundaries are not only preserved, but also remain consistent across frames, regardless of the varying environmental conditions. The architecture's robustness to these changes ensures that even as new elements or disturbances appear, the model still provides coherent and unified segmentation results, reflecting its strong capacity to maintain accuracy over time.

These success cases underline the TVSS architecture's ability to offer consistent and continuous segmentation of objects, crucial for maintaining visual coherence across video sequences. The model's strength lies in its ability to handle the temporal aspect of visual data, ensuring that segmentation evolves seamlessly across frames without disruptions.

C.2. Failure Cases

While the proposed TVSS architecture demonstrates robust performance across various scenarios, it is not without limitations. Fig. S2 illustrates two primary challenging scenarios where the model encounters difficulties.

(a) Transparent objects: The first set of failure cases involves the presence of transparent objects. Transparent materials often present ambiguous visual cues, making it challenging for segmentation models to accurately delineate boundaries and classify regions. In these instances, TVSS may misinterpret the transparency, leading to incorrect segmentation of the object or its background.

(b) Error propagation from initial mis-classification: The second set of challenges pertains to the propagation of initial segmentation errors. When the model makes an initial misclassification in a frame, this error can propagate to subsequent frames due to the temporal state-sharing mechanism. Although TVSS is designed to leverage temporal information to enhance consistency, early mistakes can temporarily degrade segmentation accuracy until corrective learning occurs in subsequent frames.

These failure cases highlight areas for potential improvement, such as incorporating specialized modules for handling transparent materials and enhancing error correction mechanisms to mitigate the impact of initial misclassifications. Addressing these challenges will further strengthen the reliability and applicability of the TVSS architecture in diverse and complex environments.

C.3. Additional Visualizations

To qualitatively analyze segmentation consistency in videos and its effect based on the number of frames used during inference, we present Fig. S3. Visual comparisons demonstrate that the results from using only one frame exhibit rough and fragmented segmentations. In contrast, predictions made using eight or thirty-two frames show smoother and more refined boundaries, closely resembling the ground truth (GT). This observation underscores the model's ability to effectively integrate temporal information, leading to better object delineation and improved segmentation boundaries. The enhanced consistency and quality of segmentation suggest that incorporating more frames enables the model to capture dynamic features and contextual information more effectively, particularly in challenging or ambiguous areas. This improvement can be attributed to the model's capacity to learn from the additional frames, resulting in a more accurate representation of the scene. This is especially apparent in complex or cluttered environments, where utilizing multiple frames significantly enhances the robustness and overall accuracy of the segmentation.