Bootstrapping Video Semantic Segmentation Model via Distillation-assisted Test-Time Adaptation

Supplementary Material

A1. Effect of Warm-Up Ratio of W2F Protocol

Figure A1 summarizes performance trends under the W2F protocol, where adaptation is restricted to the first $r\%$ of each video frames and the adapted model is thereafter fixed. Ours (DiTTA) consistently surpasses both the frame-wise inference (ISS baseline [59]) and the video-supervised VSS model (CFFM [45]), implying that our distillation-assisted adaptation yields strong temporal coherence even without requiring supervised VSS training. Compared with the combination of ISS model and the conventional TTA method (CoTTA [55]), DiTTA not only delivers higher absolute mIoU at every warm-up ratio, but also achieves a larger improvement rate as the ratio increases, indicating that our method more effectively converts additional warm-up data into segmentation quality. In particular, DiTTA exhibits notably larger performance gains as the warm-up ratio r increases, compared to other methods. Combined with the substantial performance from our full-video adaptation protocol (Table 5 in the main paper), this trend suggests that the proposed method can serve as a flexible and efficient solution adaptable to diverse deployment constraints and user requirements.

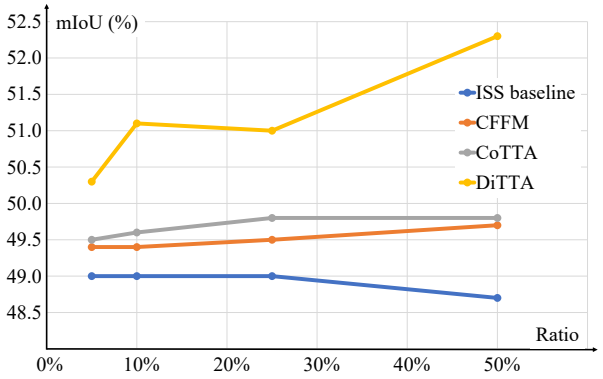

Figure A1. mIoU (%) under the W2F protocol with varying warm-up ratios. DiTTA (ours) achieves consistently higher performance and larger gains with increased ratios compared to ISS baseline, VSS (CFFM), and CoTTA.

A2. Details on Distillation Target

This section outlines the procedure for generating distillation targets in DiTTA, which is briefly described in the main paper. The goal is to combine the semantic predictions of the ISS model with the temporal consistency of SAM2, yielding temporally coherent masks that serve as supervision signals during TTA. The process consists of three stages: (1) object prompt sampling, (2) bidirectional mask propagation, and (3) mask-class assignment.

High-confidence pixels are first selected from the ISS predictions to define O object prompts. Using these prompts, SAM2 performs bidirectional propagation to generate consistent object masks across the video, effectively transferring its temporal knowledge into the ISS model during adaptation. The resulting binary masks for frame t are denoted as $M_t = \{m_t^1, m_t^2, \ldots, m_t^O\}$. Semantic labels are then assigned to each object via a scoring-based selection process.

A2.1. Sampling Object Prompts

The process begins by sampling confident pixel locations from the ISS model's predictions. For each class, the most confident pixel is selected as a candidate prompt for object, ensuring semantic diversity across objects. To avoid unreliable regions, only pixels above a confidence threshold τ are considered.

The sampled object prompts are subsequently used in the bidirectional propagation step to obtain temporally consistent masks throughout the video. As the video progresses, new object candidates may emerge due to motion or occlusion. To handle these, the sampling process is conducted in parallel with the bidirectional propagation stage, allowing dynamic capture of the newly appearing objects across frames.

A2.2. Bidirectional Propagation

Given the sampled object prompts, we leverage SAM2 to propagate the corresponding masks across the video in both forward and backward directions. Figure A2 illustrates this procedure.

Specifically, the top row of Fig. A2 shows the test video, which serves as the input for bidirectional propagation. The colored points (light blue and orange) represent examples of sampled object prompts. The second and third rows illustrate the forward propagation and backward propagation processes, respectively.

Notably, this example provides an intuitive explanation of why backward propagation is essential. In the forward propagation process, the object mask of the monitor is immediately sampled in the first frame and is propagated until the monitor moves out of the frame. In contrast, the door first appears in the second frame but is not sampled as an object prompt until the fourth frame due to the low reliability of the ISS model. Consequently, forward propagation

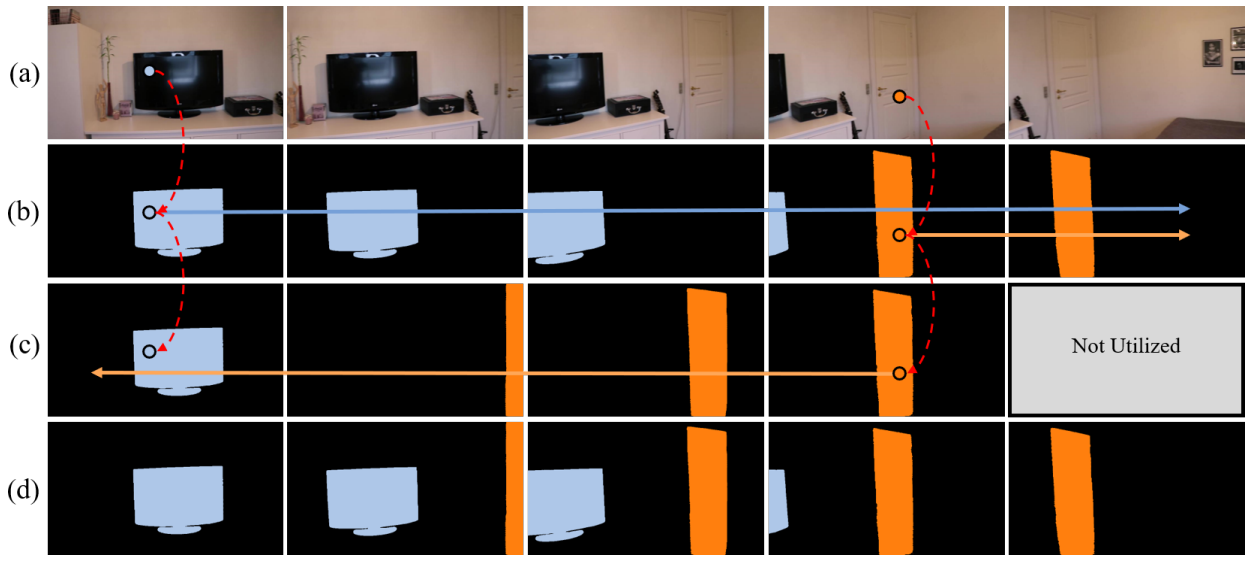

Figure A2. Visualization of spatiotemporal object mask propagation process described in the main paper. (a): Object prompt sampling, (b): Forward propagation, (c): Backward propagation, and (d): the final merged spatiotemporal object mask.

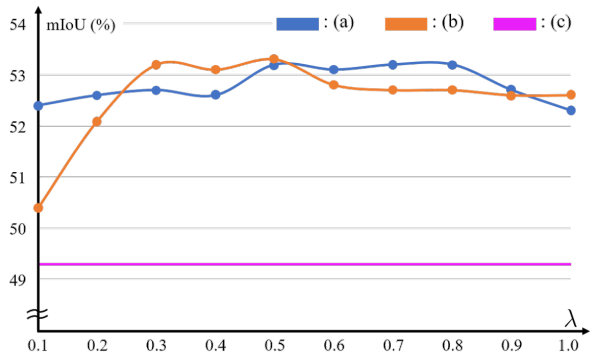

Figure A3. Verification of our semantic label assignment method. (a): $\lambda_{\text{area}} = 0.3$ while λ_{freq} varies, (b): $\lambda_{\text{freq}} = 0.8$ while λ_{area} varies, and (c): zero-shot refinement.

only provides an object mask for the door starting from the fourth frame. Backward propagation compensates for this overlook by propagating backward from the fourth frame to successfully generate the spatiotemporal mask for the door in the third and second frames. In practice, such misses are quite common during the object sampling process, making forward propagation insufficient on its own and necessitating backward propagation for correction. Finally, masks from forward and backward propagation are merged to produce the final output, as shown in the last row, which is then utilized in subsequent processes.

A2.3. Mask Class Selection

Semantic labels for each object mask are determined through a mask-class selection process that leverages the semantic predictions of the ISS model. While SAM2 provides temporally consistent object masks, it lacks semantic understanding. To compensate for this, we utilize the frame-wise predictions of the ISS model as the primary source of semantic information. In particular, the ISS outputs guide class assignment by evaluating prediction confidence, spatial dominance, and class-level statistics across the video.

To incorporate the overall semantic context of the video, we define the score α^c as in Equ. 4 in main paper. In this section, we provide a detailed explanation of each component involved. Specifically, we define $M_{t,c} = M_t \odot \mathbf{1}(\mathbf{Y}_t = c)$, which is a binary mask indicating the sub-region of the mask M_t in frame t where the ISS model predicts class c.

The default component, α^c_{rel}, represents the reliability of the ISS model's frame-wise inference for class c within a given spatiotemporal object mask region. Specifically, this is calculated as the average reliability (refer to Equ. 2) of the ISS model's predictions over all pixels within $M_{t,c}$ as:

$$\alpha^c_{\text{rel}} = \frac{\sum_{t=1}^{T}(M_{t,c} \odot \mathbf{R}_t)}{\sum_{t=1}^{T} M_{t,c}}. \tag{A1}$$

As α^c_{rel} increases, the class c is more confidently assigned.

Next, we describe the auxiliary terms supporting α^c_{rel}. γ^c_{area} represents the proportion of the mask area for object that is predicted as class c. It is computed by dividing the number of pixels in the mask regions predicted as class c by the total number of pixels in the mask, as follows:

$$\gamma^c_{\text{area}} = \frac{\sum_{t=1}^{T} M_{t,c}}{\sum_{t=1}^{T} M_t}. \tag{A2}$$

As the proportion of area predicted as class c increases, the component γ^c_{area} becomes larger. This effectively complements the first component by preventing cases where a small incorrectly predicted region with high confidence assigns the wrong class to the whole mask.

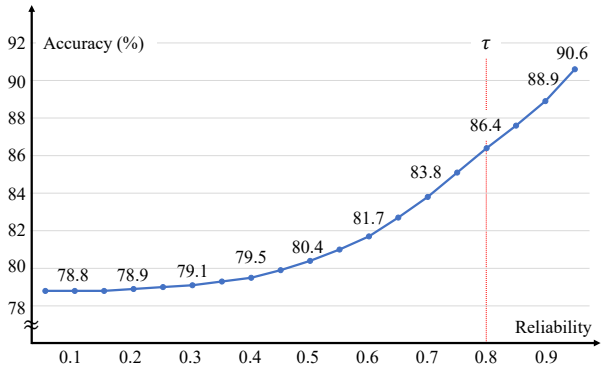

Figure A4. Reliability - Accuracy Curve in VSPW *val* set.

The final component γ^c_{freq} represents the overall distribution of class c across the entire video frames, according to the ISS model predictions. It is calculated as the proportion of pixels predicted as class c in the entire frames relative to the total number of pixels in the video:

$$\gamma^c_{\text{freq}} = 1 - \frac{\sum_{t=1}^{T} \mathbf{1}(\mathbf{Y}_t = c)}{T \cdot H \cdot W}. \tag{A3}$$

Since the ratio is subtracted from 1, this component decreases as the proportion of regions predicted as class c across the video increases, helping adjust for long-tail classes that appear infrequently.

Finally, the score α^c is computed by integrating all three components, enabling class assignment for the distillation targets while accounting for ISS prediction reliability, the class distribution within masks, and the class imbalance.

$$\alpha^c = \alpha^c_{\text{rel}} \cdot (\gamma^c_{\text{area}})^{\lambda_{\text{area}}} \cdot (\gamma^c_{\text{freq}})^{\lambda_{\text{freq}}}, \tag{A4}$$

To demonstrate the robustness of our method, we evaluate the effect of the hyperparameters λ_{area} and λ_{freq}. Figure A3 summarizes all results across λ ablation. In (a), λ_{area} was fixed at 0.3, while λ_{freq} was varied from 0.1 to 1.0 in increments of 0.1. In (b), λ_{freq} was fixed at 0.8, while λ_{area} was adjusted in the same manner with (a). Lastly, (c) shows the results of the zero-shot refinement, aforementioned in the main paper. In all configurations, (a) and (b) consistently outperformed (c) by a significant margin, demonstrating robust performance across most parameter ranges. These results highlight both the effectiveness and robustness of our method.

A3. Analysis on Reliability

In obtaining distillation targets, we utilize entropy-based reliability as defined in Equ. 3. This follows the common observation from prior works such as TENT [53] and CoTTA [55], which assume that predictions with lower entropy are more likely to be correct. Figure A4 shows the relationship between the reliability of the pre-trained ISS model and the per-pixel classification accuracy on the VSPW validation set. Consistent with earlier studies, we observe a clear positive correlation: pixel-wise accuracy increases with reliability.

In particular, when the reliability value exceeds 0.8, the accuracy reaches approximately 90%, indicating a high level of trustworthiness. We further validate this observation by conducting an ablation study under the full-video adaptation setting, varying the reliability threshold τ around 0.8. The best performance was achieved at $\tau = 0.8$, supporting its effectiveness as a practical criterion. Based on these findings, we adopt 0.8 as the reliability threshold, treating predictions above this value as sufficiently trustworthy for distillation.

Table A1. Class mapping from VSPW [37] to Cityscape [7] for cross-dataset experiments. We merge the *person* and *rider* classes in Cityscape, as both correspond to the *person* class in VSPW. Additionally, the *traffic sign* class in Cityscape is excluded from evaluation due to the lack of a corresponding class in VSPW.

VSPW [37]	Cityscapes [7]
road	road
path	sidewalk
building	building
wall	wall
handrail, fence	fence
pole	pole
traffic light	traffic light
-	traffic sign
wood, tree, flower, other plant	vegetation
grass	terrain
sky	sky
person	person, rider
car	car
truck	truck
bus	bus
train	train
motorcycle	motorcycle
bicycle	bicycle

A4. Details on Cross-Domain Experiments

This section provides additional details about the cross-domain adaptation experiments described in Cross-dataset Scenario section of the main paper. We adapt the ISS model (SegFormer [59] with the MiT-B5 backbone) pre-trained on VSPW [37] to each video sample in the validation set of Cityscapes [7]. For quantitative evaluation on Cityscapes, we use frames with GT semantic segmentation annotations, which are available for one frame per video, following the protocol of prior works.

As VSPW and Cityscapes datasets differ in their class definitions, we define a mapping between similar semantic classes from VSPW to Cityscapes, as shown in Table A1. Since the VSPW dataset does not include the *rider* class or

a closely related equivalent, we merge the *rider* and *person* classes in Cityscapes for evaluation. We map the *person* class in VSPW to this merged person-related class. The *traffic sign* class in Cityscapes is excluded as there is no corresponding class in VSPW. According to the class mapping, the logits from the pre-trained ISS model are re-defined for the Cityscapes dataset by aggregating only the logits corresponding to VSPW classes. Classes in VSPW that are not included in the mapping are ignored during aggregation. For multi-to-one mappings, such as from *wood*, *tree*, *flower*, and *other plant* (VSPW) to *vegetation* (Cityscapes), we take the maximum logit among the relevant VSPW classes.

In addition, we conduct experiments on ADE20K [70] to VSPW. Similar to the VSPW-to-Cityscapes experiments, we identify overlapping classes between the datasets and perform evaluations only on those classes. We employ a ISS model checkpoint pretrained on ADE20K and directly apply DiTTA to it.

A5. More Implementation Details

This section provides additional details on our implementation. To implement mask-based contrastive alignment, we adopt the concept of a memory bank. A memory bank list is maintained for prototype feature vectors and updated at every iteration. This design enables stable training while effectively incorporating temporal consistency. All experiments are conducted using a single RTX 3090 GPU.

A6. More Qualitative Comparisons

We provide more qualitative comparisons for the experiments conducted in the main paper. Specifically, Fig. A5 and Fig. A6 compare the results of DiTTA under W2F protocol, where the ratio is set to 10%., such as CoTTA [55] or AuxAdapt [41]. Furthermore, Fig. A7 and Fig. A8 show cross-dataset results from VSPW [37] to Cityscapes [7] and from ADE20K [70] to VSPW, respectively. In addition to the results in the main paper, these substantial results clearly demonstrate the superiority and practicality of the proposed DiTTA.

Figure A5. Additional qualitative comparison of VSS results across different methods under W2F protocol. Only the initial 10% of frames are used for adaptation, with evaluation conducted on the remaining frames. (a): Frames, (b): ISS baseline, (c): AuxAdapt, (d): CoTTA, (e): Zero-shot Refinement, (f): DiTTA, and (g): GT.

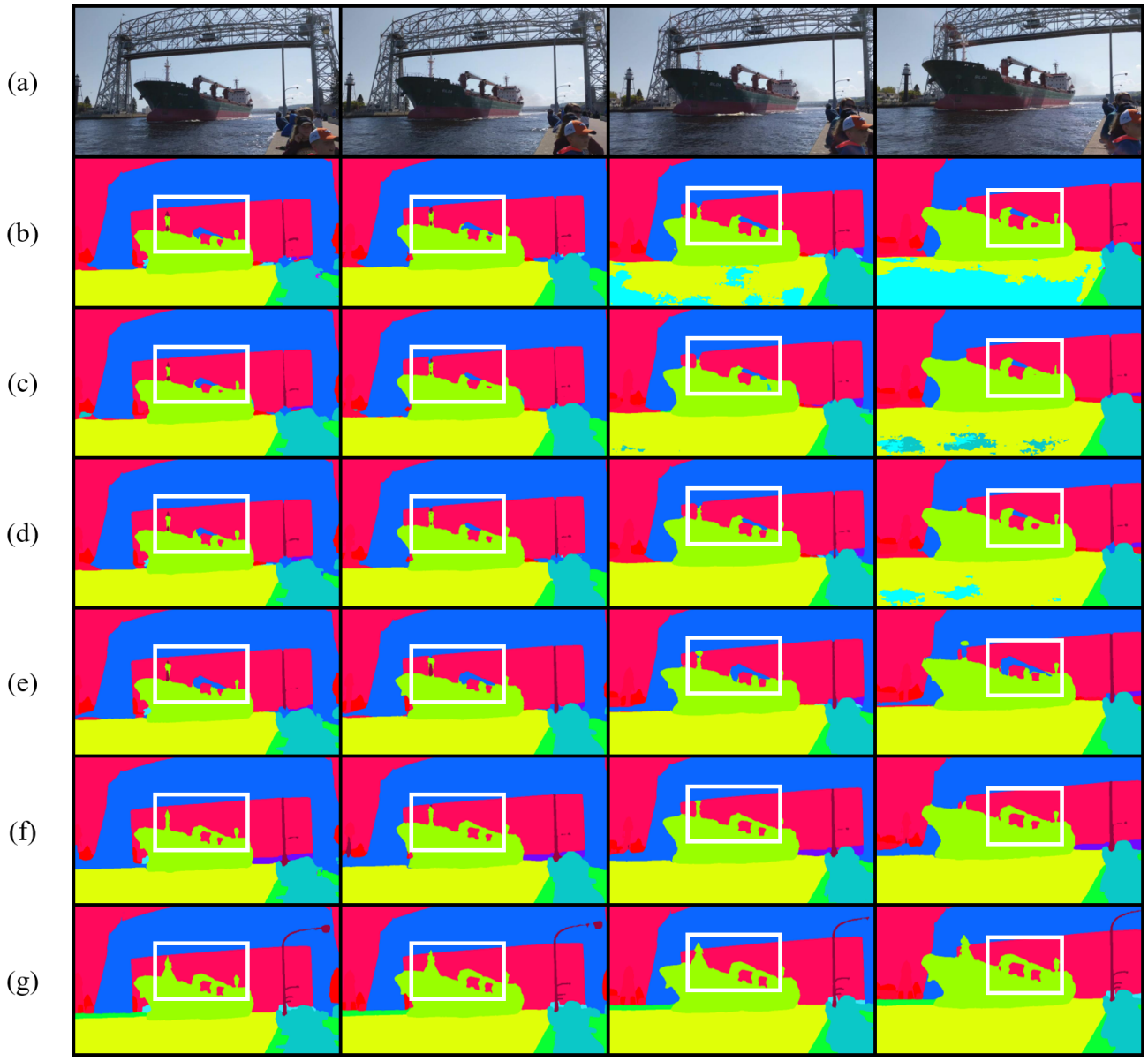

Figure A6. Additional qualitative comparison of VSS results across different methods under W2F protocol. Only the initial 10% of frames are used for adaptation, with evaluation conducted on the remaining frames. (a): Frames, (b): ISS baseline, (c): AuxAdapt, (d): CoTTA, (e): Zero-shot Refinement, (f): DiTTA, and (g): GT.

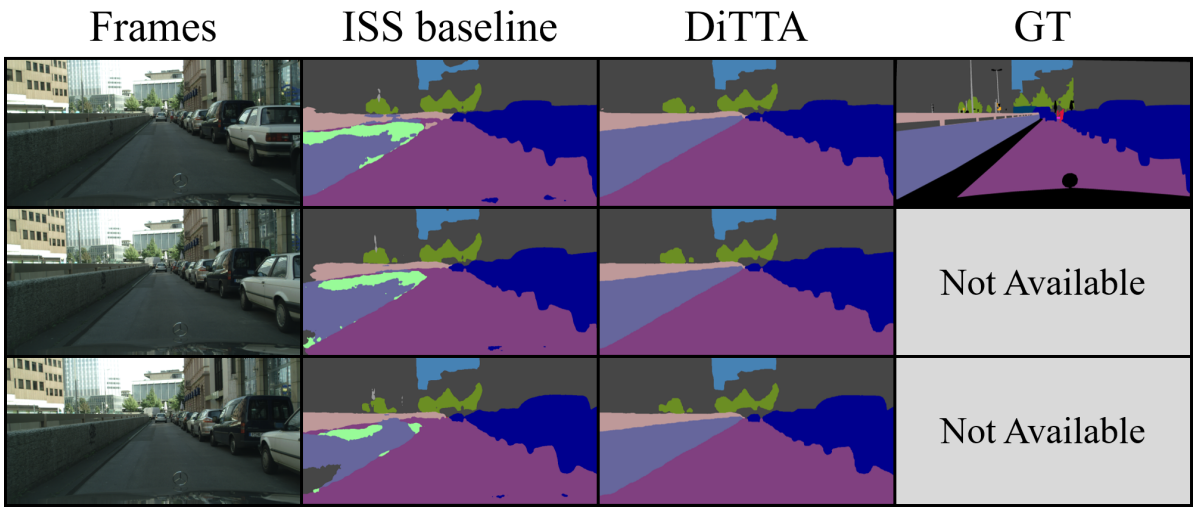

Figure A7. Qualitative comparison regarding the cross-domain experiments from VSPW to Cityscape (under W2F protocol).

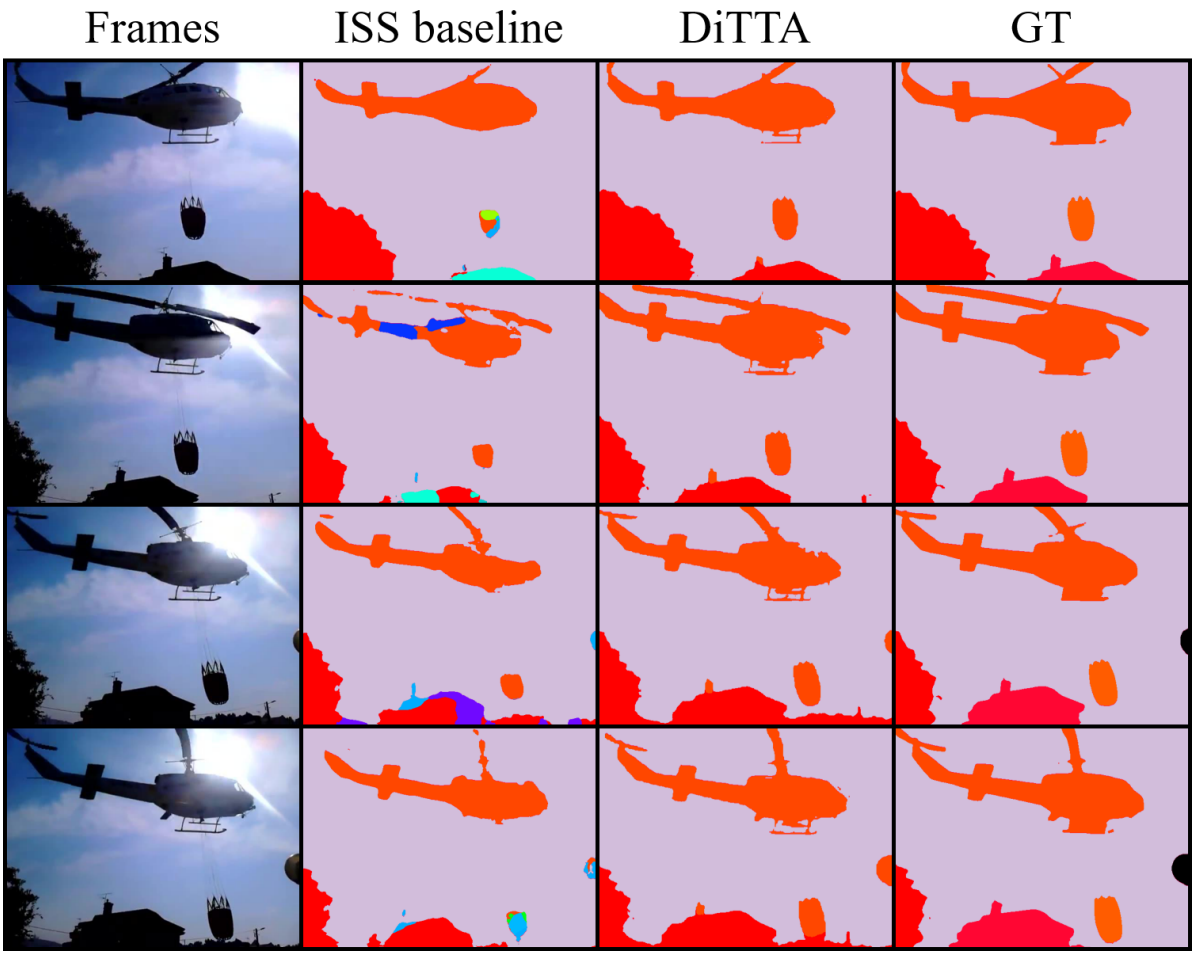

Figure A8. Qualitative comparison regarding the cross-domain experiments from ADE20K to VSPW (under W2F protocol).