—Supplementary Material—
Video-kMaX: A Simple Unified Approach for Online and Near-Online Video Panoptic Segmentation

Inkyu Shin[1,2†] Dahun Kim[2] Qihang Yu[2†] Jun Xie[2] Hong-Seok Kim[2] Bradley Green[2]
In So Kweon[1] Kuk-Jin Yoon[1] Liang-Chieh Chen[2†]
[1]KAIST [2]Google Research

Appendices

In this supplementary material, we provide

A) Comparison with Normal Cross-Attention (Sec. A)

B) Analysis on Memory Matching Space (Sec. B)

C) Analysis on Model Complexity (Sec. C)

D) Algorithm for our LA-MB (Sec. D)

E) Visualization Analysis (Sec. E)
- More qualitative results
- Structural prior learned by queries
- Failure case & Limitation

A. Comparison with Normal Cross-Attention

As discussed in the Sec. 3.1 of main paper, we deliberately design our clip-kMaX with the k-means cross-attention [7], which we empirically found to be very effective for handling the extremely large sequence of spatially and temporally flattened clip features. We now elaborate on the experiments and particularly compare with the normal (*i.e*., original) cross-attention [3] as well as the advanced latent memory cross-attention [1] (*i.e*., the cross-attention mechanism used in TubeFormer [1], which adopts latent memory to facilitate attention learning between video frames).

Tab. 1 summarizes our findings. To ensure the fairness, we employ the same backbone Axial-ResNet50-B1 [4] that has been pretrained on ImageNet-1K and Cityscapes. The baseline, employing the normal cross-attention module, yields the performance of 68.4% STQ. The performance can be further improved by 1.6% STQ, if we adopt the latent memory [1] in the cross-attention module. By contrast, our clip-kMaX, adopting the k-means cross-attention mechanism, attains 73.9% STQ, significantly outperforming the conventional cross-attention and latent memory cross-attention by **+5.5%** and **+3.9%** STQ, respectively.

backbone	method	STQ
Axial-ResNet50-B1	normal cross-attention (baseline)	68.4
	latent memory cross-attention	70.0
	k-means cross-attention (clip-kMaX)	73.9
	+ LA-MB	74.7

Table 1. **Comparison with Normal-Cross Attention.** The k-means cross-attention adopted by our proposed clip-kMaX achieves the best STQ than the normal cross-attention and latent memory cross-attention, demonstrating the effectiveness of k-means cross-attention in video understanding task.

memory	size of matching space **S** (*i.e*., $M \times N$)			
	$\tau = 1000$		$\tau = best$ (naïve-MB: $\tau = 1$, LA-MB: $\tau = 10$)	
	average	max	average	max
naïve-MB	67.1	336	25.1	196
LA-MB	19.7	94	2.8	24

Table 2. **Quantitative analysis on matching space size** between naïve-MB and our LA-MB. The size of matching space **S** could help us understand the difficulty of matching M objects in the memory with the detected N objects in the current frame. τ is the hyper-parameter to refresh out the old objects. We consider two cases, where $\tau = 1000$ to mimic the case where we barely remove the old objects, and $\tau = best$ uses the best hyper-parameter value for each setting.

The improvement is attributed to the effectiveness of k-means cross-attention that performs the cluster-wise argmax on cluster centers. Additionally, we show that our proposed LA-MB is complementary to clip-kMaX, which sets the best STQ performance (74.7 STQ). Our results suggest that using k-means cross-attention can reduce the ambiguity in cross-attention between queries and large flattened clip features, resulting in a higher quality of video panoptic segmentation results.

B. Analysis on Memory Matching Space

In the Sec. 3.2 of main paper, we address the limitations of the previous memory buffer approach [6], referred as

method	backbone	frame size	scenario	params	FLOPs	FPS	
						average	worst
Video-kMaX	ResNet50	375×1242 (KITTI-STEP)	Online (T=1)	56.4M	83G	41.2	40.8
			Near-online (T=2)	56.4M	167G	23.4	23.3
		720p (VIPSeg)	Online (T=1)	56.4M	162G	24.4	24.0
			Near-online (T=2)	56.4M	324G	13.6	13.5
Video-kMaX	Axial-ResNet50-B1	375×1242 (KITTI-STEP)	Online (T=1)	74.2M	118G	31.0	30.8
			Near-online (T=2)	74.2M	236G	17.2	17.0
		720p (VIPSeg)	Online (T=1)	74.2M	231G	18.2	18.2
			Near-online (T=2)	74.2M	461G	9.9	9.8
Video-kMaX	ConvNeXt-L	375×1242 (KITTI-STEP)	Online (T=1)	231.6M	370G	12.9	12.8
			Near-online (T=2)	231.6M	740G	6.9	6.8
		720p (VIPSeg)	Online (T=1)	231.6M	715G	7.0	6.9
			Near-online (T=2)	231.6M	1431G	3.7	3.7

Table 3. **Model complexity**. We report the inference complexity of our Video-kMaX in terms of params, FLOPs, and FPS (frames per second) on a V100 GPU, under both the online and near-online scenarios. We report three backbones, including ResNet50, Axial-ResNet50-B1, and ConvNeXt-L, on both the KITTI-STEP and VIPSeg datasets.

naïve-MB. One of the limitations of naïve-MB is the huge matching space in memory decoding, which increases the difficulty of matching and thus results in low association quality. From that perspective, we empirically prove that our hierarchical matching scheme, LA-MB, can effectively reduce the matching space size as shown in Tab. 2. To do so, we calculate the size of the similarity matrix $\mathbf{S}$ (*i.e.*, $M \times N$, where there are M objects in the memory and N detected objects in the current frame) to quantitatively measure the matching space size. We note that modern approaches [6] adopt a memory refreshing strategy, where the old objects stored in the memory will be removed if they are τ-frame older than the current frame, which, to some degree, alleviates the issue of large matching space. However, we will show that using the memory refreshing strategy alone is not sufficient to reduce the matching space size. We compare the matching space between naïve-MB and our LA-MB under two cases of τ, which is the hyper-parameter to refresh out the old objects in the memory, affecting the matching space size. In the first case, we set τ to 1000, which mimics the ideal scenario where we have a very large memory and the old objects are barely removed, aiming to exclude the effect of refreshing strategy and focus on the memory buffer approach itself. As shown in the table, we can observe that LA-MB can greatly improve the matching space efficiency by a healthy margin (*i.e.*, 3.4× smaller and 3.6× smaller in average and max values, respectively). In the second case, τ is set to be the optimal value for each memory buffer approach (*i.e.*, 1 for naïve-MB and 10 for LA-MB). As shown in the table, the memory refreshing strategy effectively reduces the matching space size for naïve-MB. However, our LA-MB still outperforms naïve-MB by achieving 9.1× and 8.2× more efficient matching space in average and max value, respectively.

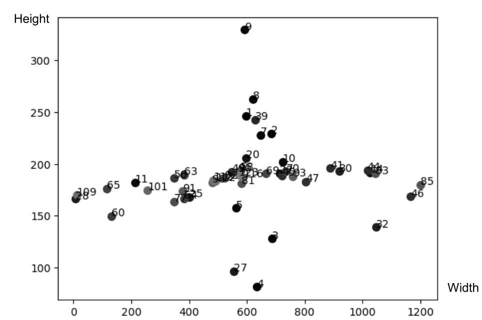


Figure 1. **Query Visualization** on KITTI-STEP *val* set. We use Video-kMaX with Axial-ResNet50-B1 backbone that is trained on KITTI-STEP and then plot the location of averaged mask center (including all stuff and things) predicted by each query.

C. Analysis on Model Complexity

In Tab. 3, we measure params, FLOPs, and FPS for our method, using a Tesla V100 GPU with CUDA 11.0 and batch size 1. We run the inference 3 times to obtain the average and worst FPS.

D. Algorithm for our LA-MB

In Alg. 1, we provide the algorithm for our **L**ocation-**A**ware **M**emory **B**uffer (LA-MB), which consists of two phases: Encoding phase to store the previous object features, and Decoding phase to associate current objects with the objects stored in the memory buffer. For better understanding, we also attach our code snippet in the supplementary materials.

Algorithm 1: Algorithm for LA-MB

Input:
1. **LA-MB** = $\{(\hat{q}_j^{t-1}, \hat{b}_j^{t-1})\}_{j=1}^{M}$ with M encoded objects until frame $t-1$.
2. Feature set (q_i, b_i) of **object** i in current frame t.
3. Panoptic map $P \in \mathbb{R}^{H \times W}$ of previous frame $t-1$

Output: (updated ID of **object** i, updated **LA-MB**)

begin
  # Decoding phase.
  if *Video-Stitch(object* i *and* P*)* $\leq M$ **then**
    # k is the updated ID of object i.
    $k \leftarrow$ Video-Stitch(object i and P)
  else
    $f(i,j) = e^{-\|b_i - \hat{b}_j\|^2/T} \cdot cos(q_i, \hat{q}_j)$
    $r = \arg\max_M (f(i,j)_{j=1}^{M})$
    if $f(i,r) \geq \alpha$ **then**
      $k \leftarrow r$
    else
      $k \leftarrow M+1$
  # Encoding phase.
  if $k \leq M$ **then**
    # The object is tracked in the memory.
    $\hat{q}_k^t = (1-\lambda)\hat{q}_k^{t-1} + \lambda q_i$
    $\hat{b}_k^t = b_i$
    for $j \in \{1, ..., k-1, k+1, ...M\}$ **do**
      $\hat{q}_j^t = \hat{q}_j^{t-1}$
      $\hat{b}_j^t = \hat{b}_j^{t-1} + (\hat{b}_j^{t-1} - \hat{b}_j^{t-2})$
  else
    # The object is new.
    $\hat{q}_k^t = q_i$
    $\hat{b}_k^t = b_i$
    for $j \in \{1, ..., M\}$ **do**
      $\hat{q}_j^t = \hat{q}_j^{t-1}$
      $\hat{b}_j^t = \hat{b}_j^{t-1} + (\hat{b}_j^{t-1} - \hat{b}_j^{t-2})$
  return (k, updated **LA-MB**)

E. Visualization Anaylsis

More qualitative results We show some visualization results in Fig. 2 for VIPSeg, where the baseline naïve-MB fails to associate persons in a crowd, since they have similar appearance features. On the other hand, our LA-MB correctly associates the same person by effectively exploiting both the appearance and location features. In our supplementary submission, we also include video panoptic segmentation results on the validation sets of KITTI-STEP [5] and VIPSeg [2]. Our Video-kMaX (consisting of clip-kMaX and LA-MB) demonstrates more clear and consistent video results than the baselines.

Structural prior learned by queries We observe that the object queries learned by our Video-kMaX demonstrate a structural prior that a particular query will respond to objects around a specific location on the image plane. To visualize the structural prior, for each query, we compute the mean location center of all its segmented objects in the whole KITTI-STEP validation set, and show the scatter plot in Fig. 1. As shown in the figure, each object query is responsible to segment objects around a specific location on the image plane. Interestingly, the object queries are scattered mostly along a vertical and a horizontal line, showing the property of ego-centric car in KITTI-STEP, where the street-view images are collected by a driving car.

Failure case and Limitation We analyze the failure mode of our Video-kMaX in Fig. 3. The first row and second row are video frames and corresponding video panoptic results with our Video-kMaX, respectively. We observe that a person initially assigned with ID number 99 until frame 2 is re-assigned with different ID numbers, *i.e*., 107 (in frame 3) and 108 (in frame 4). The ID switch could be attributed to two reasons. First, the appearance feature of the occluded person (*i.e*., person ID 107 in frame 3) is not reliable, as most of its discriminative appearance regions are occluded. Second, the target object demonstrates a large random movement, violating our slow linear motion assumption encoded by the location feature. This failure case presents a challenging but interesting research direction to further improve our model by strengthening both appearance and location features.

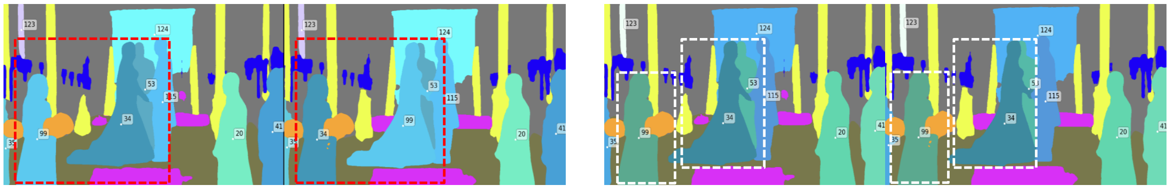
References

[1] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13914–13924, 2022. 1

[2] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 1

[4] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2020. 1



(a) naive-MB (b) LA-MB

Figure 2. **Visualization results** on VIPSeg *val* set. The baseline naïve-MB, only exploiting the appearance feature, fails to associate the same person, as neighboring people have similar appearance features. On the other hand, our LA-MB, exploiting both appearance and location features, successfully associates the same person.

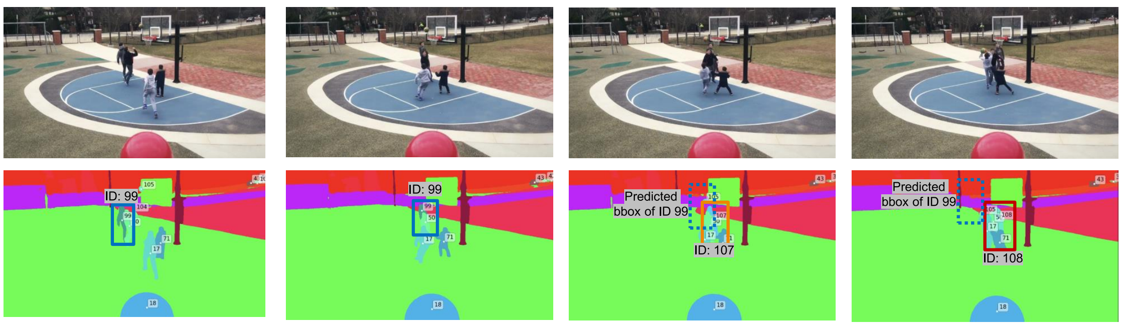


Figure 3. **Failure case** on VIPSeg *val* set. The target object is initially assigned with ID 99. Its ID switches to 107 and 108 in frame 3 and frame 4, respectively. Our method fails to track the target object, because it is heavily occluded and moves at a large random pace, making both appearance and location features unreliable.

[5] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljosa Osep, Laura Leal-Taixe, and Liang-Chieh Chen. Step: Segmenting and tracking every pixel. *Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021. 3

[6] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 588–605. Springer, 2022. 1, 2

[7] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means Mask Transformer. In *Proceedings of the European Conference on Computer Vision*, pages 288–307. Springer, 2022. 1