Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion Supplementary Material

1. Notations

This is a quick index of the notations in the order of appearance.

General notations	
r	The current interaction round index.
t^r	The index of the current user-interacting frame.
M^r	The mask results of all the frames of the current round.
M^r_j	The mask result of the j-th frame of the current round.
M^{r-1}	The mask results of all the frames of the previous round.
Notations related to interaction	
$M^{r-1}_{t_r}$	The previous mask result of the current user-interacting frame.
Notations related to propagation	
H, W	The spatial dimensions of the features after the encoder with stride 16.
C^k, C^v	The channel dimensions of the "key" and "value" features respectively.
T	The number of frames in the memory bank.
$\mathbf{k}^M, \mathbf{v}^M$	The extracted "key" and "value" features from the memory encoder.
$\mathbf{k}^Q, \mathbf{v}^Q$	The extracted "key" and "value" features from the query encoder.
$\mathbf{F}_{ij}$	The dot product between the query feature at position i and the memory feature at position j.
$\mathbf{W}_{ij}$	The normalized affinity between the query feature at position i and the memory feature at position j.
k	A hyperparameter. Top-k filtering is applied along the memory dimension.
$\text{Top}^k_j(\mathbf{F})$	Indices of affinities that are top-k in the j-th column (i.e., memory) of $\mathbf{F}$.
$\mathbf{m}_j$	The aggregated memory feature for the query position j.
Notations related to fusion	
t_i	The target frame index to be fused.
t^c	The closest previously interacted frame index in the direction of propagation.
$M^{r'}$	The current mask results after propagation and before fusion.
$\mathcal{D}^+, \mathcal{D}^-$	The positive and negative mask differences respectively.
$\mathcal{A}^+, \mathcal{A}^-$	The positive and negative mask differences aligned with the target frame t_i.
n_r, n_c	The normalized temporal distance between the target frame t_i and the current/previously interacted frames t^r and t^c respectively.

2. Comparisons with KMN

KMN [1] presents two major ideas: 1) hide-and-seek training augmentation and 2) kernelized memory reading. We focus only on 2) here. KMN assumes that each *memory* position should only attend to a local window (specified by a Gaussian

distribution with a fixed σ) in the *query*. Thus, filtering is performed on the *query* for every *memory* position. In contrast, our top-k filtering does not assume any spatial prior and performs on the *memory* for every *query* position. Ours is discrete, leading to a efficient algorithm while KMN slows the algorithm down. Note that the two methods are not at odds with each other – we can use both at the same time to obtain higher accuracy in the `test-dev` set (see our project page) without retraining.

References

[1] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, 2020. 1