Thanks for all reviewers' valuable comments. We will first answer the common questions then respond to each reviewer.

[CQ1] Sub-group sampling (R1-Q3, R3-Q1): Following the common testing protocol as adopted in Ref [44], we sequentially divide each input group into sub-groups consisting of 5 images in a non-overlapping manner. For the last sub-group with images less than 5, we supplement by randomly selecting samples from the whole given group.

[CQ2] Model size & VGG16 Backbone (R1-Q4, R3-Q4): The performance of our method with VGG16 backbone is shown in the table. 1) Our method can still achieve better performance than Ref [44]. 2) The model size of Ours-V is comparable to the method [44] (121 MB vs 119 MB). Since most CoSOD competitors did not release codes, here we only report the model size of [44] for comparison, which is provided directly by its authors. Our code will be released.

	Cosal2015				CoSOD3k				MSRC				iCoseg			
	AP	F_β	MAE	S_m	AP	F_β	MAE	S_m	AP	F_β	MAE	S_m	AP	F_β	MAE	S_m
Ref [44]	0.8846	0.8666	0.0791	0.8433	0.8245	0.8066	0.0916	0.7983	0.8217	0.7903	0.2072	0.6768	0.8979	0.8823	0.0773	0.8606
Ours-V	0.8862	0.8748	0.0644	0.8612	0.8263	0.8249	0.0696	0.8368	0.8752	0.8597	0.1139	0.8082	0.9177	0.8940	0.0416	0.8839

[CQ3] Additional parameters (R2-Q4, R4-Q1): Sorry for the unclear description. In fact, the baseline (with the ResNet50 backbone) is carefully designed to share a close number of parameters with the full model (178 MB vs 176 MB), which is adequate for proving the superiority of our CoADNet without introducing additional parameters.

[R1-Q1] Video-SOD. The input images in CoSOD task are not necessarily temporally-related, which deviates from Video-SOD that emphasizes temporal modelling. Hence, direct adaptation might not be applicable.

[R1-Q2] Effectiveness of OIaSG. Sorry for making the confusion. As mentioned in the ablation study, we have pre-trained the baseline for the SOD task on the DUTS dataset, which could prove the superiority of our OIaSG scheme.

[R2-Q1] Unclear motivation. 1) In the introduction, we have separately highlighted the three main motivations (please see Page 2, Line 41-62), which illustrate the necessity of the GASA, GGD, and GCPD modules item by item. We will make clearer statements for your concerns. 2) Our overall aggregation-and-distribution architecture for the problem of CoSOD is novel and brings very competitive performances. The GASA brings new insights in solving order-sensitivity and capturing long-range inter-image dependencies. Moreover, the GGD and GCPD further investigate group-individual interaction and co-saliency consistency that are very crucial but completely ignored in previous CoSOD methods.

[R2-Q2] Missing related works (RW). Due to limited space, we only analyzed the highly-related works [32,35,43,44]. Experiments included the most recent SOTA works for comparisons. We will add a RW section in the full version.

[R2-Q3] Feature visualization. The learned co-saliency features highlight the common and salient objects in each image, and suppress others. As visualized in Fig. 3, the features in the encoder show much higher response around co-salient objects with reduced background redundancy. We will further provide more visualizations for each module.

[R2-Q5] Weak relevance. This paper deals with CoSOD task under the NeurIPS track of *Applications -> CV*. Moreover, there have been some visual saliency researches on very recent NeurIPS's publications (e.g., [R1][R2]).

[R3-Q2] Parameter selection & block-wise group shuffling. 1) In practice, we tested several choices and found $B = 8$ works best. Actually, our model is not sensitive to B within a reasonable range. We will discuss this parameter in the ablation study. 2) As depicted in Fig. 2, for the input N images, we first split each feature map along channel axis into B blocks, and concatenate all the N blocks coming from the same b^{th} partition.

[R3-Q3] Order-insensitivity. Order-sensitivity is caused by the sequential channel concatenation of individual features. In GASA, we apply channel-wise softmax to each shuffled features that are composed of several blocks, and then make element-wise summation of these blocks. In GCPD, we assemble the individual feature vectors and similarly apply softmax across channels and make summation. The two modified feature combination methods are order-invariant.

[R4-Q2] Inconsistancy of [44] and VGG16 results: The reported results in [44] adopts the VGG16 backbone. In our experiments, we tested the results of [44] provided by the authors, in which HRNet [R3] is used as backbone and hence causes inconsistency (our reported results are better). Although HRNet [R3] is stronger than VGG16 and ResNet50, our model (with VGG or ResNet backbone) still achieves superior performance. Please see **[CQ2]** for the VGG16 results .

[R4-Q3] Idea of group semantics. Our solution only shares a similar big picture with [32] in terms of aggregating group semantics. However, this paper explored new insights under a two-step aggregation-and-distribution framework. Instead of directly duplicating and concatenating the group semantics with individuals, we designed GGD for dynamic group-individual combination and suppression of distracting information redundancy, which turns to be very crucial but is ignored in previous studies. Besides, the GASA differs from [32] in attentive learning and long-range modelling.

[R4-Q4] AP. We list AP comparisons of [44] and ours in **[CQ2]**. We will report APs for all methods in the final version.

[R4-Q5] Saliency priors. In the CoSOD task, maintaining awareness of salient regions and knowing how to exploit saliency priors for co-saliency mining are critical. Compared with common practice of SOD pretraining, our OIaSG provides a more effective and flexible jointly-optimized workflow for integrating more reliable saliency guidance information, which is the first attempt for CoSOD. Ablation study also supports this.

References: [R1] M. Zhang, *et al.*, Memory-oriented decoder for light field salient object detection. *NeurIPS*, 2019.
[R2] T. Nguyen, *et al.*, DeepUSPS: Deep robust unsupervised saliency prediction via self-supervision. *NeurIPS*, 2019.
[R3] K. Sun, *et al.*, Deep high-resolution representation learning for human pose estimation. *CVPR*, 2019.