We thank all the reviewers for their constructive comments, which are addressed as follows.

To Reviewer 1:

Q1: Move experiments from supp. to the main paper. Thanks for this suggestion, and we will move the LVIS instance segmentation results and COCO panoptic segmentation results from supplementary to the main paper.

Q2: Missing citations. Thanks for your suggestion. The suggested citations will be added.

To Reviewer 2:

Q1: Comparisons with other dynamic operations [1, 2]. In video object segmentation, Yang *et al*.[2] apply conditional *batch normalization* to manipulate the intermediate feature maps, making the feature map focus on a specific object instance. Similarly, AdaptIS [1] predicts the affine parameters, which scale and shift the features conditioned on each instance. It requires a large mask head to achieve good results. Both the operations used in [1, 2] belong to the more general scale-and-shift operation, which can roughly be seen as an attention mechanism on intermediate feature maps. In contrast, the dynamic convolution naturally suits our purpose of dynamic mask representation, as it directly decouples the object mask generation into mask kernel prediction and mask feature learning. Quantitatively, we compared with [1] in Table 3 of supplementary, where DFIS shows 6.2 PQ and 9.3 PQ^{things} better than [1]. We should add the above discussions to the related work.

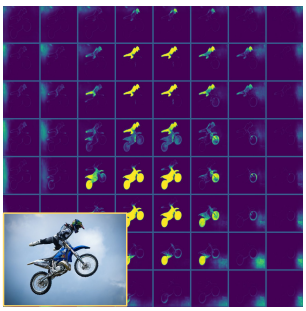

Figure A: Position-aware property. The duplicated masks will be suppressed by the proposed Matrix NMS.

Q2: Visualization of position-aware property in DFIS. We show an example in Figure A. More visualization will be provided.

Q3: Comparison with concurrent work CondInst [3]. The dynamic scheme part is somewhat similar, as they both are inspired by Dynamic Filter Networks [Brabandere et al. NIPS'16]. But the methodology is different. (a) [3] relies on the relative position to distinguish instances as in AdaptIS, while DFIS uses absolute positions as in SOLO. It means that [3] needs to encode the position information N times for N instances, while DFIS performs it all at once using the global coordinates, regardless how many instances there are. (b) [3] works in a two-stage manner that first performs box detection and then segments the detected objects. In contrast, our proposed DFIS is much simpler, which takes an image as input, directly outputs instance masks and corresponding class probabilities. For example, [3] has at least 4 loss terms while DFIS has 2 loss terms. (c) In addition, our Matrix NMS serves as an important contribution.

To Reviewer 3:

Q1: Comparisons with AdaptIS and CondInst. We address the concerns of the comparisons with AdaptIS [1] and the concurrent work CondInst [3] in R2.Q1 and R2.Q3.

Q2: The details of coordinate channels. A tensor of the same spatial size as input is created, which contains pixel coordinates normalized to $[-1, 1]$. We then concatenate it to the input features and feed to the following layers. Specifically, given the $H \times W \times D$ shaped feature tensor, the size of the new tensor is $H \times W \times (D + 2)$, where the last two channels are x-y pixel coordinates.

Q3: The contribution of dynamic mask representation in terms of speed. Using the same NMS strategy, the Res-101 DFIS runs at 15.1 FPS on a V100 GPU vs. SOLO's 11.6 FPS. The speed-up mostly comes from fewer convs (7 to 4), as the dynamic mask representation enables us to achieve superior results with lighter prediction head.

To Reviewer 4:

Q1: About the claim of 'no need to predict the bounding box'. We agree that the bounding box serves as an effective and efficient representation for object localization. We are not saying that BBox object detection is not important. The point is that our instance segmentation method doesn't reply on BBox prediction. We will remove "...instance segmentation should be an advance alternative..." in L65 to eliminate the confusion.

Q2: The relationship between DFIS and bottom-up instance segmentation. We agree that the bottom-up method is a long-standing research line and has shown competitive results on some datasets, *e.g.*, Cityscapes. But the key difference here is that the bottom-up method needs post-processing to group pixels into individual object masks, while DFIS directly predicts explicit object masks, without the restriction of pixel grouping.

Q3: Others. L65 will be removed. About the L79, the comparison is on methodology. The quantitative comparison including real-time efficiency is shown in Figure 1(a) and Table 1. About L194, thanks for pointing it out. s_i and s_j are the confidence scores. All above explanations will be added to our paper to make it more clear.