

Supplementary Materials: Unveiling and Mitigating Bias in Audio Visual Segmentation

Anonymous Authors

A PRELIMINARY OBSERVATION

In order to control necessary variables and obtain more accurate statistical results, more effort in observing the following variables during the experiment has been made.

A.1 Area of target region

Both traditional segmentation and existing audio-visual segmentation methods often disregard small objects and display a preference towards objects of a larger area, which is a prevalent factor impacting the mIoU.

To demonstrate the impact of this issue, we also conducted a statistical analysis on the influence of object area on mIoU within AVS. It is found that the proportion of object area to the total image area also affects mIoU.

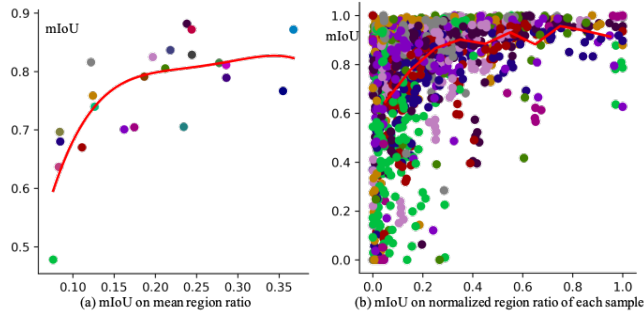


Figure A1: The impact on mIoU on the area of the target object in V1S. The red line is the fitted curve over the data points. Each point in Figure (a) corresponds to a class of objects and its average area ratio, while each point in Figure (b) corresponds to a video sample and its area ratio with normalization. In Figure (a), we calculate the mean area of each semantics and observe its overall impact on mIoU. On the sample level, in Figure (b), by conducting statistical analysis on the area and mIoU of each sample, we found that the phenomenon of a larger area usually results in higher mIoU. Therefore, controlling the proportion of the object area is essential for analysis.

Therefore, when assessing the influence of audio semantics on mIoU, it is necessary to scale the original images. So we scaled each target object to occupy 30% to 40% of the overall image for observation and experimentation in Figure 2 of the main paper.

A.2 Number of instances per sample

In traditional segmentation, it is commonly believed that the complexity will bring difficulty to the model. It is not clear about the complexity. For a better understanding of this in AVS, observations have been made to prove that both the number of target semantics

and target instances in a video sample will cause variance in the performance of segmentation in Figure A2.

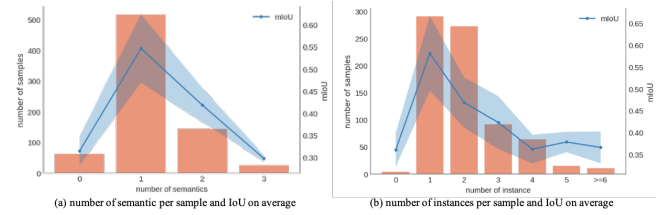


Figure A2: The impact on mIoU by the number of target semantics or target instances in a video sample in V1M. In the graph, the blue curve represents the fluctuation of the mean value, while the blue shade represents the variance. The orange color indicates the number of samples. (a) The increase in the number of target semantics in a video tends to result in a decrease in mIoU. (b) The increase in the number of target instances in a video tends to result in a decrease in mIoU.

Therefore, when constructing the Co-AVS dataset, our goal was to demonstrate the cooperative capability over semantics. To control variables, we specifically selected scenarios where there are only two semantic categories, and each category has only one instance, to build the dataset.

B IMPLEMENTATION DETAIL

B.1 Avoid permanent dormant queries

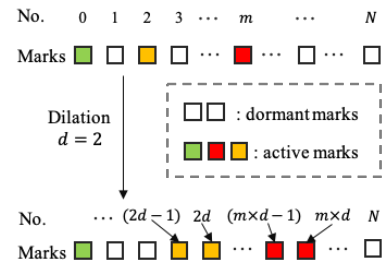


Figure B3: The illustration of simple dilation taking $d=2$ as an example.

The learnable queries function to generate a sufficient set of potential regions to acquire suitable regions. However, our active query strategy has resulted in a reduction number of effective proposed regions, consequently decreasing the number of valid proposals. To address this issue, we introduce the simple dilation

method in Fig. B3 with a dilation factor of d , enabling one latent semantic marker to align with multiple active queries. This approach guarantees that the mask of the same semantic is proposed by the same set of specified queries while providing flexibility of focus on any one of the multiple active queries.

B.2 Minor adjustment between binary and semantic segmentation

The debias strategy for semantic segmentation reorganizes the mask and class logits distribution separately. However, during the implementation of binary segmentation, the operation of the mask and class logits is equivalent to the reorganization of only the mask logits without background. Further detail about logits reorganization is discussed in Sec.D.3.

B.3 Curriculum learning for AVSS

During the training process of AVSS, we employed a curriculum learning training strategy that involved using V1S, V2 (samples not in V1S and V1M), and V1M subsets in sequential order from easy to difficult.

C ABLATION OBSERVATION

C.1 Mask generated by each query

In previous AVS models that used per-mask segmentation, each query can generate masks with various semantics, as shown in the first row of the figure. However, this approach is unable to accommodate the additional tasks of perception and interaction that the transformer decoder needs to handle in the AVS task.

As mentioned in Sec.3.1 of the paper, we aim to assign specific semantic perception tasks to each learnable query in order to further regularize semantic interactions within the transformer decoder to the designated queries. The observation on each active query is shown in Figure C4. From Ex.1 to Ex.5, we can observe the semantic similarities between the masks generated by same query, and the segmentation results are also intact. This demonstrates that the model is able to follow and update the semantic information we allocate. Moreover, each mask result is intact and satisfactory, which provides qualitative supplementary evidence for Figure 5 in the paper, further confirming the effectiveness of using learnable queries to capture latent audio corporations.

In summary, although the masks generated by active queries may be partially influenced by the perception module, we believe that the coarse filtering is sufficient to improve model performance. Additionally, a small number of anomalous semantic masks do not necessarily lead to errors of categories in semantic segmentation, as the result categories depend on the class logits. Therefore, a small number of anomalous semantic masks do not hinder the overall effectiveness.

D DISCUSSION

D.1 Possible Versatility on Ref-VOS

What's also interesting is that we conducted experiments on Referring Video Object Segmentation (Ref-VOS) [1] without validating the existence of visual priors. Similar to AVS, Ref-VOS, as a Cross-Model Guided Video Segmentation approach, also achieved

a limited increase of 0.9% in the score (the product of mIoU and Fscore).

Table D1: Possible versatility on Ref-VOS. Both AVS and Ref-VOS are a task of cross-model guided segmentation. “↑” signifies a positive effect achieved by employing the contrastive debiased strategy compared to the vanilla method, while “↓” indicates a negative effect.

Ref-VOS	Backbone	Method	†Score
Referformer [1]	Swin	None	56.2
Referformer [1]	Swin	Debias (Logits ensemble)	55.7 ↓
Referformer [1]	Swin	Debias (Bias-Only)	56.3 ↑
Referformer [1]	Swin	Debias (Uncertainty-based)	57.1 ↑

† Score: The scores mentioned here are reproduced by us. Due to our limited GPU resources, we made modifications to Referformer to accommodate these constraints, resulting in lower overall performance than the original paper. However, the comparisons between different versions still remain representative.

D.2 Contrastive of the audio of other semantic

Suppose we directly use audio with random semantics to replace the mute or noise-only guidance in the biased branch in Sec.4.4 (like using violin sound instead of mute or noise-only audio to get the visual prior in the bus video). In that case, certain negative effect occurs on bias handling. Without any audio semantic guidance, the results can fully reflect the visual bias. However, using audio cues with random semantics introduces redundant **audio semantic information** rather than just **visual prior** to the logits distribution, often failing to reflect pure visual bias. In the experiment, we also attempted to use random audio guidance as guidance in the biased branch, but the performance in addressing the bias suffered from a decrease from 0.8% to 2.1% on V1M using the contrastive debias method mentioned in Sec.4.4. So, the contrastive learning by constructing positive and negative audio-visual pairs does not apply to the purpose of bias handling here.

D.3 Debias strategy operation on logits

The debias strategy for AVSS (semantic segmentation) reorganizes both mask logits and class logits, but when treated as binary segmentation, it is equivalent to reorganizing only the mask logits. It makes us wonder about which reorganization is actually functioning, mask logits, class logits, or both. To investigate the actual mechanisms of our debias strategy, we conducted further exploration.

When performing **per-mask** semantic segmentation, there are obviously four operation schemes: reorganize ① mask logits, ② class logits, ③ mask and class logits separately, and ④ overall logits (multiplication of mask and class logits). If we simplify the basic debias strategy to directly subtracting the logits, it is important to note that from the perspective of matrix multiplication, ③ and ④ are clearly not equivalent as

$$\begin{aligned}
 & (logits_{mask} - logits_{mask}^b) \times (logits_{class} - logits_{class}^b) \\
 & \neq logits_{mask} \times logits_{class} - logits_{mask}^b \times logits_{class}^b. \quad (1)
 \end{aligned}$$

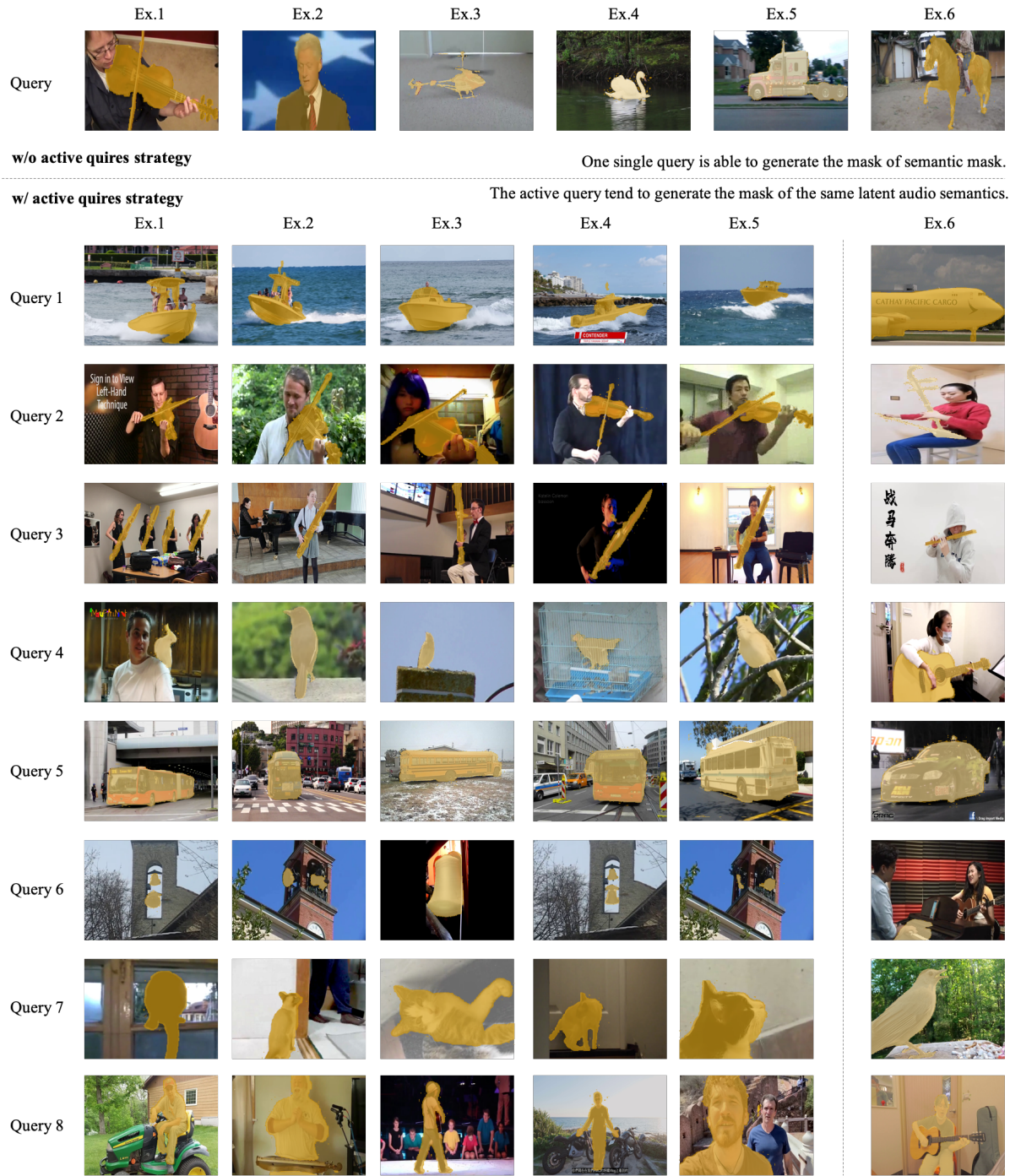


Figure C4: The binary mask generated by active queries in both single-source and multi-source audio scenery. The masks in each row are generated by the same learnable query. In the perception process, we assign semantic information to specific learnable queries as active queries. From Ex.1 to Ex.5, we can observe the semantic similarities between the masks, and the segmentation results are also intact. This demonstrates that the model is able to follow and update the semantic information we allocate. Additionally, since we obtain latent semantic information through the audio perception module, there may be some interference from similarities in sounds produced by different objects. Therefore, it is inevitable that the model will also aggregate a small number of similar-sounding semantics. For example, the bell sound and piano sound with Query 6, and the bird sound and cat sound with Query 7.

while $logits_{mask}$ and $logits_{class}$ are the logits from vanilla audio-visual method, $logits_{mask}^b$ and $logits_{class}^b$ are the logits from biased branch.

In our empirical result, we evaluated all four methods and determined that ③ produced the most favorable outcomes. However, ④ exhibited a decrease of 1.4%. This can be attributed to the per-mask training approach, which inherently optimizes class and mask independently. So, in per mask segmentation, debiasing must be operated on each type of logit individually.

However, when it comes to versatility experiments focusing on **per-pixel** segmentation, there is no such thing as class logits and mask logits. So, for all other per-pixel segmentation methods, overall logits are reorganized based on the debias strategy.

REFERENCES

- [1] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. 2022. Language as Queries for Referring Video Object Segmentation. arXiv:2201.00487 [cs.CV]