

Supplementary Materials: SelM: Selective Mechanism based Audio-Visual Segmentation

Anonymous Author(s)

This supplementary material mainly includes two aspects:

- (1) More details on DAM and BCSM (Figure 1 and Figure 2).
- (2) More experimental results.
 - Qualitative comparisons on S4 setting (Figure 3);
 - Ablations on AVSS setting (Table 1);
 - Ablations on fusion stages (Table 2);
 - Ablations on auxiliary loss coefficient (Table 3);
 - Parameters about SelM (Table 4);
 - Ablation study on Learnable Query Dimensions (Table 5);
 - Various initialization techniques (Table 6);
 - Ablation study on DAM under AVSS setting (Table 7);
 - Ablation study on BCSM under AVSS setting (Table 8).

```
def DAM_A2VBlock(x, # Visual Feature [B T C H W]
                 y # Audio Feature [B T C 1]
                 ):
    x = reshape(x) # [B T H*W C]
    x = Dropout(Gelu(Conv1d(x))) # align
    y = Conv1d(y) # align

    # compute sim_map(visual as query, audio as key and value)
    sim_map = cross_attention(x, y) # [B T 1 H*W 1]

    # get the weighted audio feature
    y = reshape(y) # [B T 1 1 C]
    y = reshape(sim_map*y) # [B T H*W C]
    y = Norm(Conv1d(y)) # align

    # conduct a multiply of weighted audio and visual Feature
    x = x*y # [B T H*W C]
    x = Dropout(Gelu(Conv1d(x))) # output the Aligned Feature

    return x
```

Figure 1: Example of DAM-A2VBlock Algorithm. This pseudocode outlines the steps for aligning audio and visual features.

Method	AVSS	
	$M_J \uparrow$	$M_F \uparrow$
w.o.DAM	31.2(-0.7)	36.4(-0.8)
w.o.BCSM	30.3(-1.6)	35.4(-1.8)
w.o.CLEVER	28.4(-3.5)	33.2(-4.0)
SelM	31.9	37.2

Table 1: Comparative Analysis of Module Performance.

```
def BCSM_A2V_Sem(x, # Visual Feature [B T C H W]
                 y # Audio Feature [B T C]
                 ):
    # pass the audio feature through the vanilla mamba block
    # with the layernorm and skip-connection
    y = y+mamba(Norm(y)) # for three layers

    # the shape of y : [B T C]
    # use linear layer for align
    # use SiLU layer for map features to [0,1]
    audio_gate = SiLU(linear(y)) # [B T C]

    # conduct a multiply for audio gate and Visual Feature
    # with skip-connection
    x = x+audio_gate*x

    return x
```

Figure 2: Illustration of the BCSM-A2V_Sem Procedure. The pseudocode details the process for enhancing audio-visual correspondence through a mamba block with layer normalization and skip connections. It includes a SiLU-activated linear layer for feature mapping, followed by an audio gating mechanism to integrate audio and visual features.

Stage	S4		MS3	
	$M_J \uparrow$	$M_F \uparrow$	$M_J \uparrow$	$M_F \uparrow$
1	81.5(-2.0)	89.9(-1.3)	55.9(-4.4)	67.1(-4.2)
1+2	81.4(-2.1)	89.7(-1.5)	56.6(-3.7)	67.9(-3.4)
1+2+3	81.6(-1.9)	89.9(-1.3)	57.0(-3.3)	67.3(-4.0)
4	81.9(-1.6)	89.8(-1.4)	57.2(-3.1)	67.8(-3.5)
4+3	81.5(-2.0)	89.9(-1.3)	57.8(-2.5)	69.3(-2.0)
4+3+2	81.6(-1.9)	89.9(-1.3)	57.3(-3.0)	67.8(-3.5)
1+2+3+4(SelM)	83.5	91.2	60.3	71.3

Table 2: Evaluation of Sequential Stages. Stages 1 to 4 represent the four stages of the video encoder and audio encoder, respectively.

Coefficient	S4		MS3	
	$M_J \uparrow$	$M_F \uparrow$	$M_J \uparrow$	$M_F \uparrow$
0.5	82.9(-0.6)	90.8(-0.4)	59.9(-0.4)	70.2(-1.1)
1(SelM)	83.5	91.2	60.3	71.3
1.5	82.7(-0.8)	90.9(-0.3)	59.6(-0.7)	70.3(-1.0)
2.0	82.8(-0.7)	90.5(-0.7)	59.9(-0.4)	70.3(-1.0)

Table 3: Ablation study on aux loss coefficient.

Method	AVSS	
	$M_J \uparrow$	$M_F \uparrow$
A2V-SeM	31.1(-0.8)	36.6(-0.6)
V2A-SeM	30.0(-1.9)	35.5(-1.7)
BCSM	31.9	37.2

Table 8: Ablation study on BCSM directions under AVSS setting.

Method	Params
DAM	4.81M
BSCM	11.31M
CLEVER	11.63M
AVSegformer	186.1M(+4.2)
SelM	181.9M

Table 4: Parameters about SelM.

Query dim	S4		MS3	
	$M_J \uparrow$	$M_F \uparrow$	$M_J \uparrow$	$M_F \uparrow$
128	83.2(-0.3)	90.5(-0.7)	60.0(-0.3)	70.3(-1.0)
256(SelM)	83.5	91.2	60.3	71.3
512	83.0(-0.5)	90.5(-0.7)	59.3(-1.0)	69.9(-1.4)

Table 5: Ablation study on Learnable Query Dimension in CLEVER.

Method	From Scratch		Pretrain on S4	
	$M_J \uparrow$	$M_F \uparrow$	$M_J \uparrow$	$M_F \uparrow$
SelM-R50	54.5	65.6	58.4(+3.9)	70.7(+5.1)
SelM-PVT	60.3	71.3	63.4(+3.1)	73.3(+2.0)

Table 6: Ablation study on Various Initialization Techniques.

Method	AVSS	
	$M_J \uparrow$	$M_F \uparrow$
A2V Block	29.9(2.0)	35.0(-2.2)
V2A Block	29.3(-2.6)	34.1(-3.1)
DAM	31.9	37.2

Table 7: Ablation study on DAM directions under AVSS setting.

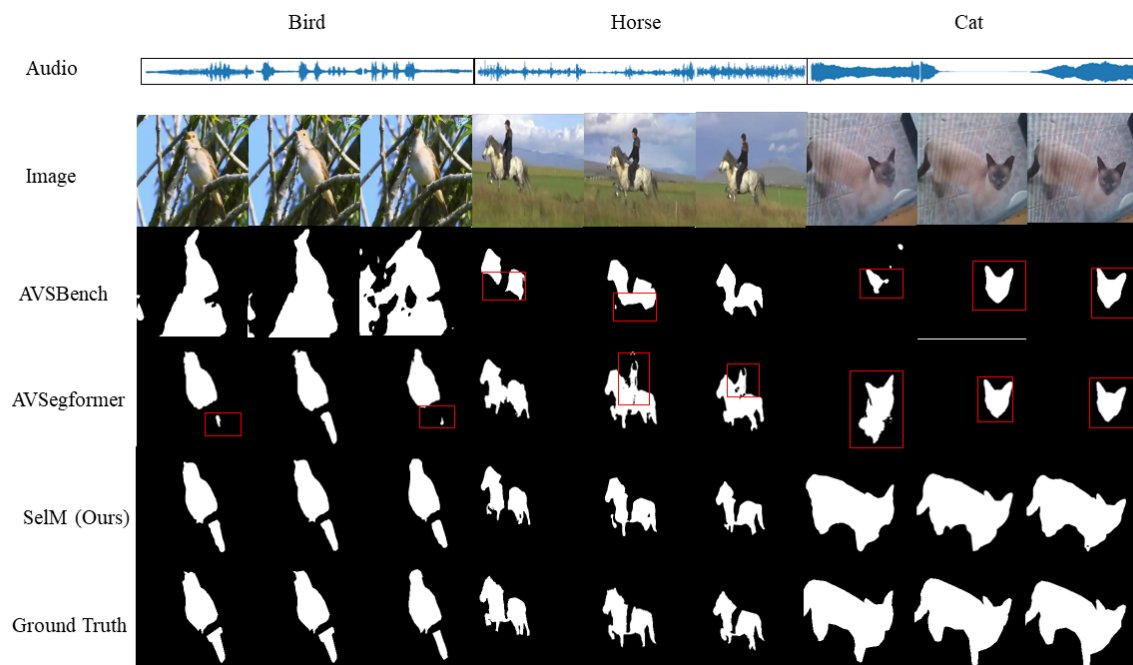


Figure 3: Qualitative Comparisons on S4 Setting.