SAVE: Speech-Aware Video Representation Learning for Video-Text Retrieval

Supplementary Material

In this supplementary material, we provide additional experiments not included in the main paper, including:
- t-SNE visualization of more audio encoders (Sec. S1)
- Using Whisper as the audio encoder (Sec. S2)
- Video-to-Text retrieval results (Sec. S3)
- Using a larger backbone (Sec. S4)
- Qualitative results (Sec. S5)

S1. Visualization of More Audio Encoders

To verify that the lack of speech semantic understanding observed in AST [9] is a general issue, we extend our analysis to ImageBind [8] and Whisper [36]. Following the same experimental setup as Fig. 1a in the main paper, we visualize the audio embeddings for *sound clips* and *narration clips* across three animal classes using t-SNE. As shown in Fig. S1, while all three models produce distinct clusters for *sound clips*, the embeddings for *narration clips* remain heavily entangled across all encoders. Notably, even Whisper, despite its robust ASR capabilities, fails to disentangle speech semantics in its audio feature space. This confirms that current audio backbones predominantly capture acoustic characteristics rather than high-level speech semantics.

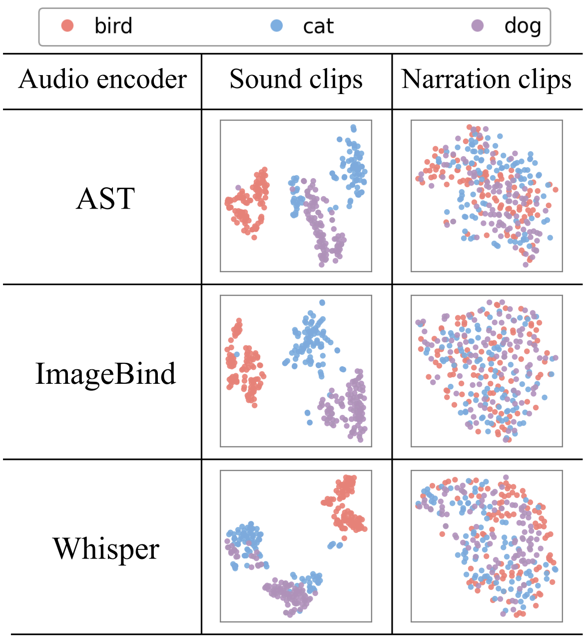

Figure S1. **t-SNE visualizations of feature spaces across three audio encoders**. All three audio encoders struggle to capture speech semantics.

S2. Using Whisper as the Audio Encoder

Tab. S1 presents an ablation study examining whether a dedicated speech branch remains necessary when using Whisper, which already possesses ASR capability, as the audio encoder. We introduce SAVE-2, which uses Whisper as the audio encoder, and SAVE-3, which further removes the speech branch. Comparing SAVE-2 and SAVE-3, we observe that removing the speech branch consistently degrades performance on both MSRVTT-9k and MSRVTT-7k, indicating that explicit speech modeling remains beneficial even when the audio encoder itself is ASR-capable.

Table S1. **Ablation study of using Whisper as the audio encoder**. Backbone: CLIP (ViT-B/32).

Setup	Audio	Speech	Soft-ALBEF	MSRVTT-9k		MSRVTT-7k	
				R1	SumR	R1	SumR
SAVE	AST	Whisper	✓	**51.3**	**216.2**	33.5	165.8
SAVE-2	Whisper	Whisper	✓	50.9	215.8	**33.8**	**166.8**
SAVE-3	Whisper	✗	✓	50.2	213.0	32.9	164.0

S3. Video-to-Text Retrieval Results

For a more comprehensive evaluation, we also report video-to-text retrieval performance on MSRVTT-9k, see Tab. S2. Same as text-to-video retrieval task (Tab. 3), SAVE achieves the highest SumR among all compared methods.

S4. Using a Larger Backbone

To verify if our method's superiority holds with a larger backbone, we repeat the main experiments using CLIP (ViT-B/16). As shown in Tab. S3, our SAVE consistently outperforms both vision-only and audiovisual baselines.

S5. Qualitative Results

To further validate the effectiveness of our speech-aware video representation for text-to-video retrieval, we present qualitative visualizations in Fig. S2. SAVE consistently retrieves the correct video because the dedicated ASR-based speech branch effectively captures speech-related cues that PIG and AVIGATE fail to exploit.

Figure S2. **Qualitative comparison** of the Top-1 retrieved videos from **PIG**, **AVIGATE**, and **SAVE**, with the same color-coded blocks indicating how each model ranks these candidate videos. Benefiting from the speech branch, **SAVE** consistently assigns the ground-truth video (with green box) the best rank, showing clear advantages on speech-related queries. Best viewed on screen.

Table S2. **The video-to-text retrieval performance of different methods**. Dataset: MSRVTT-9k. Backbone: ViT-B/32.

Model	R1	R5	R10	SumR
Vision-only:				
CLIP4Clip, Neucom22 [32]	43.1	70.5	81.2	194.8
X-Pool, CVPR22 [10]	44.4	73.3	84.0	201.7
TS2-Net, ECCV22 [30]	45.3	74.1	83.7	203.1
X-CLIP, MM22 [33]	46.8	73.3	84.0	204.1
DiCoSA, IJCAI23 [20]	46.7	75.2	84.3	206.2
PromptSwitch, ICCV23 [3]	46.0	74.3	84.8	205.1
UATVR, ICCV23 [4]	46.9	73.8	83.8	204.5
UCoFiA, ICCV23 [48]	47.1	74.3	83.0	204.4
ProST, ICCV23 [26]	46.3	74.2	83.2	203.7
DGL, AAAI24 [52]	43.5	70.5	80.7	194.7
EERCF, AAAI24 [44]	44.7	74.2	83.9	202.8
TempMe, ICLR25 [42]	45.6	72.4	81.2	199.2
DiscoVLA, CVPR25 [41]	47.7	73.6	83.6	204.9
Audiovisual:				
EclipSE, ECCV22 [27]	44.7	71.3	82.8	198.8
TEFAL, ICCV23 [17]	47.1	75.1	84.9	207.1
AVIGATE, CVPR25 [18]	**49.7**	75.3	83.7	208.7
AVIGATE+	49.5	75.5	85.4	210.4
SAVE	48.9	**78.0**	**86.5**	**213.4**
SAVE-h	46.7	76.0	85.6	208.3

Table S3. **The retrieval performance of different methods on CLIP ViT-B/16 backbone**. Dataset: MSRVTT-9k.

Model	**Text-to-Video Retrieval**				**Video-to-Text Retrieval**			
	R1	R5	R10	SumR	R1	R5	R10	SumR
Vision-only:								
CLIP4Clip, Neucom22 [32]	46.4	72.1	82.0	200.5	45.4	73.4	82.4	201.2
X-Pool, CVPR22 [10]	48.2	73.7	82.6	204.5	46.4	73.9	84.1	204.4
TS2-Net, ECCV22 [30]	49.4	75.6	85.3	210.3	46.6	75.9	84.9	207.4
X-CLIP, MM22 [33]	49.3	75.8	84.8	209.9	48.9	76.3	85.4	210.6
STAN, CVPR23 [28]	50.0	75.2	84.1	209.3	-	-	-	-
UATVR, ICCV23 [4]	50.8	76.3	85.5	212.6	48.1	76.3	85.4	209.8
UCoFiA, ICCV23 [48]	49.8	74.6	83.5	207.9	49.1	77.0	83.8	209.9
ProST, ICCV23 [26]	49.5	75.0	84.0	208.5	48.0	75.9	85.2	209.1
DGL, AAAI24 [52]	48.3	71.8	80.6	200.7	45.7	74.0	82.9	202.6
EERCF, AAAI24 [44]	49.9	76.5	84.2	210.6	47.8	75.3	85.4	207.3
TeachCLIP, CVPR24 [45]	48.0	75.9	83.5	207.4	-	-	-	-
TempMe, ICLR25 [42]	49.0	74.4	83.3	206.7	47.6	75.3	85.4	208.3
DiscoVLA, CVPR25 [41]	50.5	75.6	83.8	209.9	49.2	76.0	84.7	209.9
PIG, ICCV25 [23]	51.2	75.1	84.5	210.8	-	-	-	-
Audiovisual:								
TEFAL, ICCV23 [17]	49.9	76.2	84.4	210.5	-	-	-	-
AVIGATE, CVPR25 [18]	52.1	76.4	85.2	213.7	51.2	77.9	86.2	215.3
SAVE	**54.5**	**80.3**	**87.4**	**222.2**	**51.4**	**80.1**	**87.0**	**218.5**