浙江大学
ZHEJIANG UNIVERSITY

Uni-Code

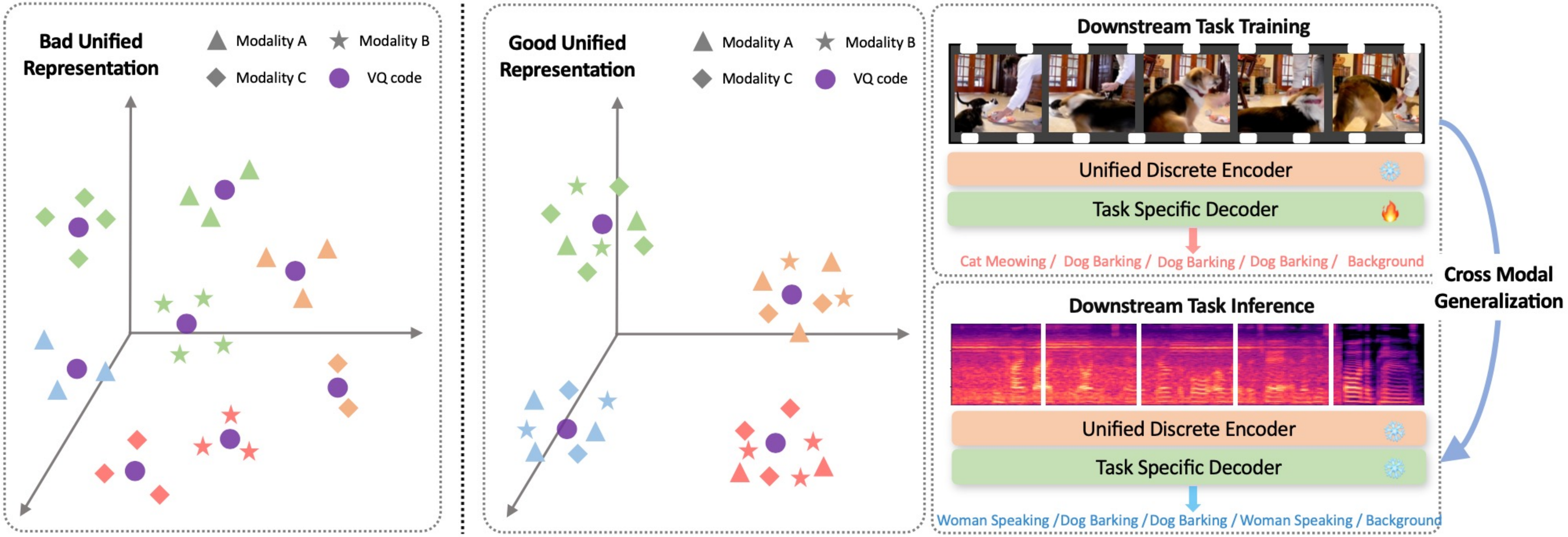
Achieving Cross Modal Generalization with Multimodal Unified Representation

Yan Xia   Hai Huang   Jieming Zhu   Zhou Zhao

01

Introduction

01 Introduction



The overview of our proposed Cross-Modal Generalization (CMG) tasks.

02

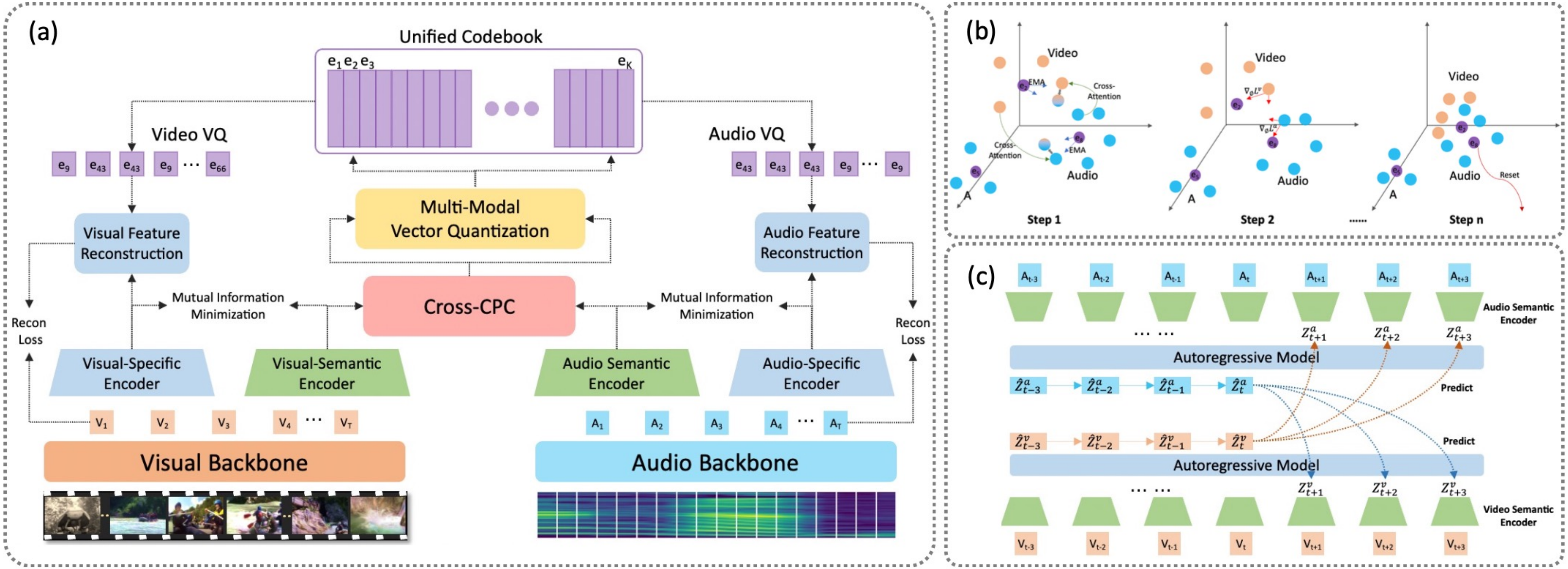
Approach

Dual Cross-modal Information Disentangling (DCID)

Multi-Modal Exponential Moving Average (MM-EMA)

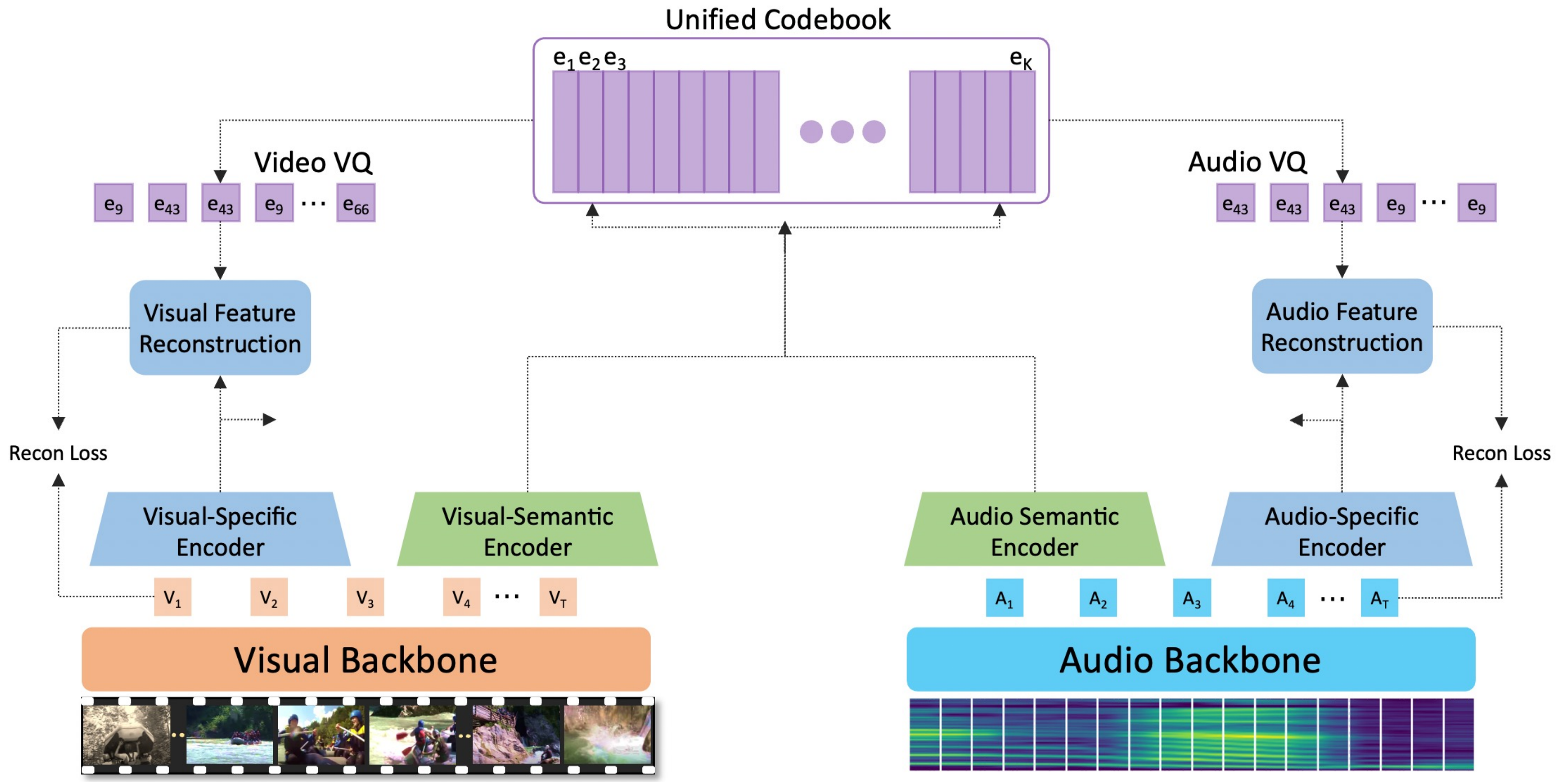
Approach

ZHEJIANG UNIVERSITY



The overview of our proposed framework: Uni-Code

Baseline：



$$\mathbf{z}_i^a = \Phi^a(\mathbf{x}_i^a) \quad \bar{\mathbf{z}}_i^a = \Psi^a(\mathbf{x}_i^a) \quad \mathbf{z}_i^b = \Phi^b(\mathbf{x}_i^b) \quad \bar{\mathbf{z}}_i^b = \Psi^b(\mathbf{x}_i^b)$$

$$L = \underbrace{\|\mathbf{x}_i^m - D(\hat{\mathbf{z}}_i^m; \bar{\mathbf{z}}_i^m)\|_2^2}_{\text{reconstruction loss}} + \underbrace{\|\text{sg}[\phi^m(\mathbf{x}_i^m)] - \mathbf{e}\|_2^2}_{\text{VQ loss}} + \underbrace{\beta\|\phi^m(\mathbf{x}_i^m) - \text{sg}[\mathbf{e}]\|_2^2}_{\text{commitment loss}}, \quad m \in \{a, b\}.$$

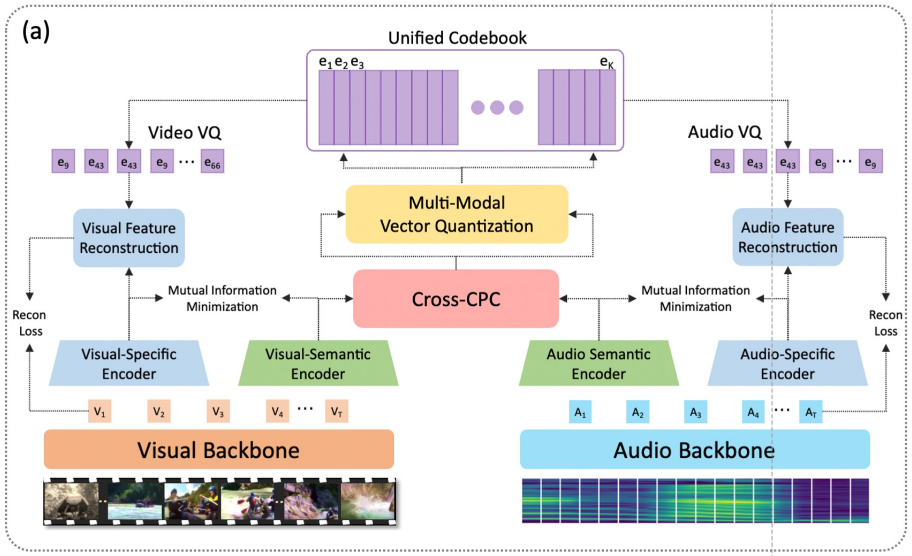
Approach: Dual Cross-modal Information Disentangling

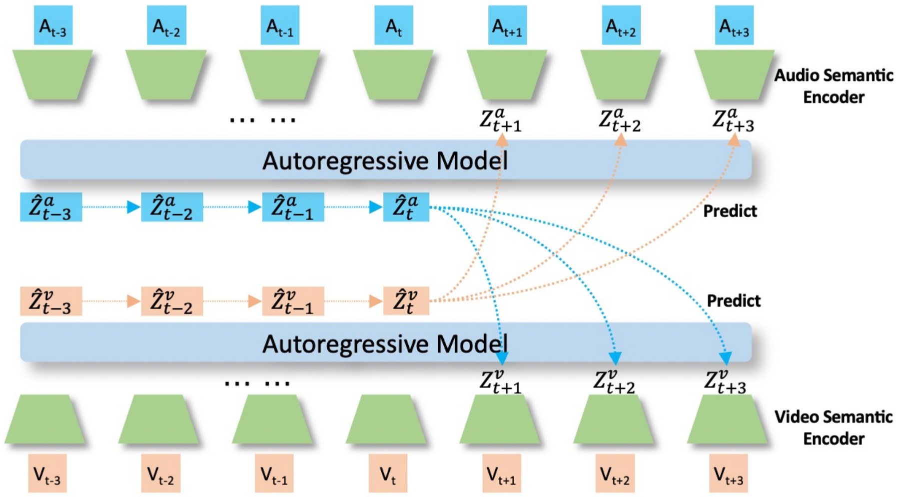
MI minimization with CLUB

$$\hat{I}_{\text{vCLUB}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T}\sum_{t=1}^{T}\log q_\theta(\bar{\mathbf{z}}_i^m|\mathbf{z}_i^m) - \frac{1}{N}\frac{1}{T}\sum_{j=1}^{N}\sum_{t=1}^{T}\log q_\theta(\bar{\mathbf{z}}_j^m|\mathbf{z}_i^m)\right], \quad m \in \{a, b\}$$

MI maximization with Cross-CPC

$$L_{cpc}^{a2b} = -\frac{1}{K}\sum_{k=1}^{K} log\left[\frac{exp(\mathbf{z}_{t+k}^b W_k^a \mathbf{c}_t^a)}{\sum_{\mathbf{z}_j \in Z_b} exp(\mathbf{z}_j^b W_k^a \mathbf{c}_t^a)}\right]; \quad L_{cpc}^{b2a} = -\frac{1}{K}\sum_{k=1}^{K} log\left[\frac{exp(\mathbf{z}_{t+k}^a W_k^b \mathbf{c}_t^b)}{\sum_{\mathbf{z}_j \in Z_a} exp(\mathbf{z}_j^a W_k^b \mathbf{c}_t^b)}\right],$$





Approach: Multi-modal Exponential Moving Average

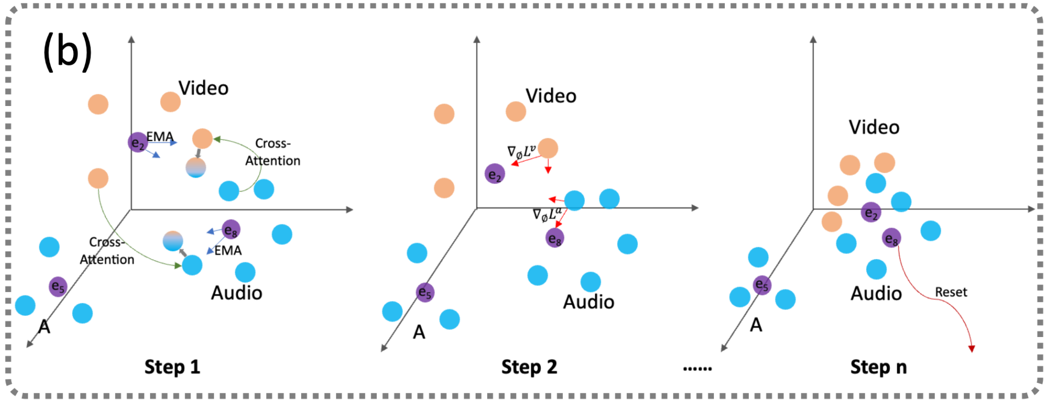
We use Cross-Attention to extract related information from the opposite modality:

$$\mathbf{r}_i^b = \text{cross-att}(\mathbf{z}_i^a; \mathbf{z}_i^b; \mathbf{z}_i^b)$$

Given a code vector $\mathbf{e}_i$, we can obtain n_i^a semantic vectors of modality A $\{\mathbf{z}_{i,j}^a\}_{j=1}^{n_i^a}$ and n_i^b semantic vectors of modality B $\{\mathbf{z}_{i,j}^b\}_{j=1}^{n_i^b}$ that are quantized to $\mathbf{e}_i$:

$$N_i^{(t)} = \gamma N_i^{(t-1)} + (1-\gamma)[n_i^{a(t)} + n_i^{b(t)}] \quad \mathbf{e}_i^{(t)} = \mathbf{o}_i^{(t)} / N_i^{(t)}$$

$$\mathbf{o}_i^{(t)} = \gamma \mathbf{o}_i^{(t-1)} + (1-\gamma)\Big[\sum_{j=1}^{n_i^{a(t)}} \frac{\mathbf{z}_{i,j}^{a(t)} + \mathbf{r}_{i,j}^{b(t)}}{2} + \sum_{j=1}^{n_i^{b(t)}} \frac{\mathbf{z}_{i,j}^{b(t)} + \mathbf{r}_{i,j}^{a(t)}}{2}\Big],$$



A new commitment loss:

$$L_{commit}^a = \beta \|\phi^a(\mathbf{x}_i^a) - \text{sg}[\mathbf{e}_i^a]\|_2^2 + \frac{\beta}{2} \|\phi^a(\mathbf{x}_i^a) - \text{sg}[\mathbf{e}_i^b]\|_2^2$$

03

Experiments

Pre-train tasks:

- Audio-Visual
- Audio-Visual-Text

Downstream tasks:

- Cross-modal event classification
- Cross-modal event localization
- Cross both modal and dataset localization/classification
- Cross-modal video segmentation
- Cross-modal retrieval

Cross-modal Event Classification & Localization Tasks：

浙江大学 ZHEJIANG UNIVERSITY

Table 1: Compared with state-of-the-art methods on two downstream tasks. We use precision to indict the performance of the models on AVE tasks, and use accuracy for AVVP tasks.

Method	VGGsounds-AVEL 24K				VGGsounds-AVEL 40K				VGGsounds-AVEL 81K			
	AVE		AVVP		AVE		AVVP		AVE		AVVP	
	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V
Baseline	4.4	5.9	7.6	8.4	5.5	5.4	6.9	8.7	7.1	9.3	5.6	7.2
S-Mit[43]	12.7	16.9	17.2	22.8	14.4	15.9	19.0	22.3	13.4	17.0	20.9	22.8
MST[13]	13.3	19.0	25.7	29.1	19.5	23.1	22.7	24.5	18.6	20.5	19.1	24.8
CODIS[16]	18.5	22.0	29.4	33.7	20.8	26.4	35.1	37.9	28.5	30.2	34.0	37.8
TURN[20]	17.7	21.0	29.4	32.4	19.1	24.3	36.9	39.3	27.6	31.4	33.8	38.1
CMCM[17]	28.9	35.9	42.6	50.4	32.7	36.8	41.9	45.1	31.1	34.0	39.3	44.8
DCID+S-Mit	28.1	32.3	45.9	49.2	32.2	34.0	47.8	53.0	34.8	37.6	51.9	53.5
DCID+MST	31.2	35.0	50.7	52.1	34.9	37.8	54.4	59.1	33.5	35.4	57.1	59.2
DCID+TURN	29.4	35.3	53.4	56.0	29.7	36.9	55.2	58.2	31.9	36.8	56.2	60.9
DCID+CODIS	33.4	36.0	53.8	60.2	36.7	41.0	52.6	62.0	35.9	40.1	54.3	59.0
DCID+CMCM	34.1	38.8	57.6	60.8	36.4	42.9	58.7	62.8	38.8	41.4	57.5	60.5
Uni-Code	**44.0**	**49.7**	**61.9**	**65.7**	**47.7**	**52.3**	**64.0**	**65.6**	**41.2**	**45.6**	**60.5**	**61.7**

Ablation Studies about our proposed modules:

浙江大学 ZHEJIANG UNIVERSITY

Table 2: Ablation studies of audio-visual pre-training on AVE and AVVP tasks.

CLUB	Cross-CPC	MM-EMA	Reset code	L_{cmcm}	VGGsounds-AVEL 24K				VGGsounds-AVEL 40K			
					AVE		AVVP		AVE		AVVP	
					V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V
-	✓	✓	✓	✓	34.9	35.1	50.6	54.0	37.2	40.3	52.9	59.5
✓	-	✓	✓	✓	4.6	5.8	10.9	24.6	5.2	7.1	12.3	24.1
-	-	✓	✓	✓	29.8	34.6	30.4	32.5	35.2	36.9	32.3	34.0
✓	✓	-	✓	✓	34.1	38.8	57.6	60.8	36.4	42.9	58.7	62.8
✓	✓	✓	-	✓	37.8	41.2	59.1	61.5	38.9	40.3	55.4	62.1
✓	✓	✓	✓	-	39.7	42.6	58.2	62.1	41.3	46.0	58.7	62.8
✓	✓	-	-	✓	28.2	30.9	41.4	49.2	31.5	33.8	46.0	48.3
✓	✓	✓	✓	✓	**44.0**	**49.7**	**61.9**	**65.7**	**47.7**	**52.3**	**64.0**	**65.6**

Table 3: Ablation studies of audio-visual-text pre-training on three downstream tasks.

CLUB	Cross-CPC	MM-EMA	Reset code	L_{cmcm}	VGGsounds-AVEL 40K							
					AVE		AVVP		AVE→AVVP		UCF(v)↔VGG(a)	
					V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V
-	✓	✓	✓	✓	50.2	51.8	62.4	66.2	50.1	51.2	9.87	9.59
✓	-	✓	✓	✓	43.8	49.2	59.3	61.1	45.5	50.6	60.6	54.6
✓	✓	-	✓	✓	52.9	49.9	62.0	67.3	48.1	46.8	66.5	60.5
✓	✓	-	-	✓	33.0	35.5	56.7	61.2	7.4	12.6	43.3	35.2
✓	✓	✓	-	✓	50.8	47.8	56.4	61.1	47.9	50.4	60.0	49.8
✓	✓	✓	✓	-	52.4	54.5	57.5	**72.9**	50.5	48.7	**69.9**	59.7
✓	✓	✓	✓	✓	**54.1**	**55.0**	**63.4**	71.0	**53.0**	**52.4**	67.1	**60.6**
Evaluation results of the labeled modality					64.8	65.8	71.0	72.9	-	-	80.0	85.4

Cross-modal video segmentation & Retrieval：

Table 4: Performance on AVS-S4 datasets (pre-trained on audio-visual-text modalities).

Methods	A2T		T2A	
	mIoU	F-score	mIoU	F-score
Baseline	69.8	81.4	69.9	81.3
Our full model	**78.0**	**87.1**	**77.7**	**86.7**
SST [49] (A2A)	60.3	80.1	-	-
AVS [48] (A2A)	78.7	87.9	-	-

Table 5: Performance of audio retrieval tasks under cross modal generalization directions.

Methods	V2T		T2V	
	R@5	R@10	R@5	R@10
Baseline	0.47	1.03	0.62	0.85
Our full model	**10.3**	**21.9**	**8.47**	**16.7**

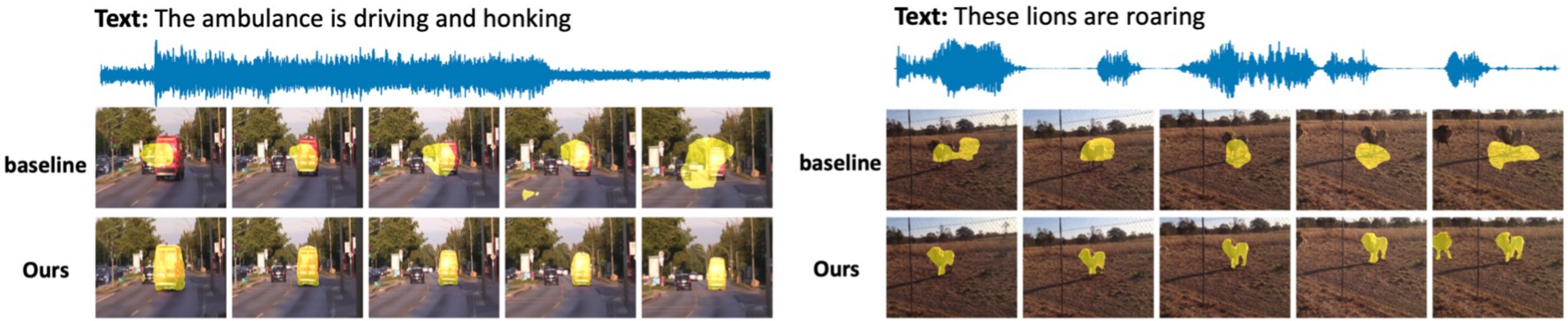


Figure 5: Visualization results of A2T (left) and T2A (right) of our model on AVS-S4 dataset. We compare our method with the baseline model.

浙江大学 ZHEJIANG UNIVERSITY

Achieving Cross Modal Generalization with Multimodal Unified Representation

Thank you

Yan Xia