Exploring the Role of Audio in Video Captioning

Yuhan Shen†, Linjie Yang*, Longyin Wen*,
Haichao Yu*, Ehsan Elhamifar†, Heng Wang*

† Northeastern University
* ByteDance

7th MUltimodal Learning and Applications Workshop (MULA 2024)

Video Captioning

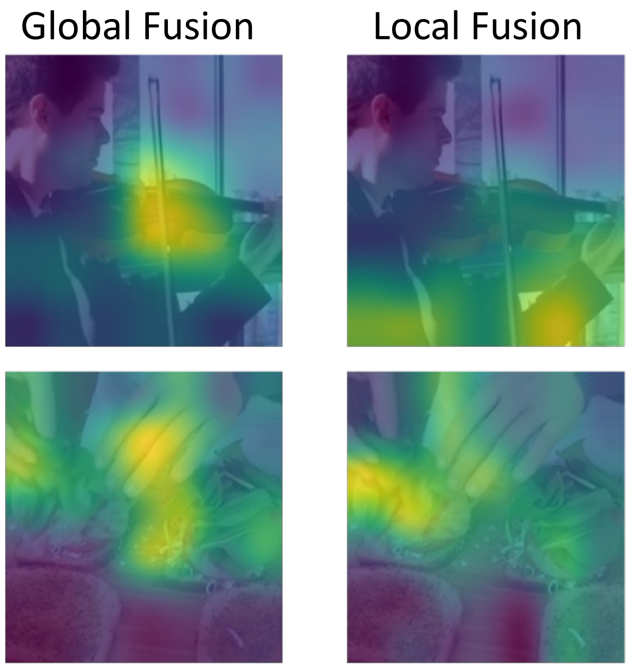
- Video Captioning: generate text descriptions of videos
- Modality: vision; audio; both
- Proposed: a pre-training framework for audio-visual video captioning

Caption: A baby fusses and cries while a woman talks and laughs.

Caption: A little girl is pointing to pictures in a book while an adult talks to her.

Challenges

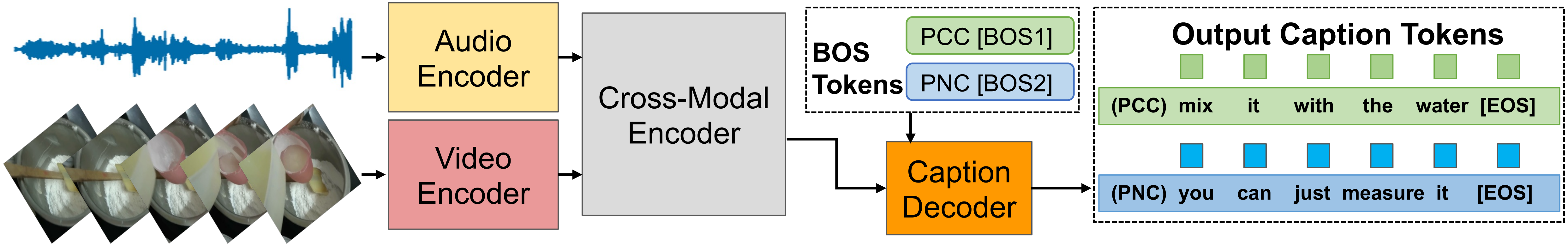
- Lack large-scale annotated datasets for video captioning pre-training
  - *Use ASR transcripts as text supervision, e.g. HowTo100M*
- ASR transcripts can be solely obtained from audio modality
  - Modality Balancing Pre-training
- Information exchange between audio and video modality
  - Local-global cross-modal fusion modules



1. A. Miech, et al. Howto100M: Learning a text-video embedding by watching hundred million narrated video clips. ICCV 2019.

Proposed Framework

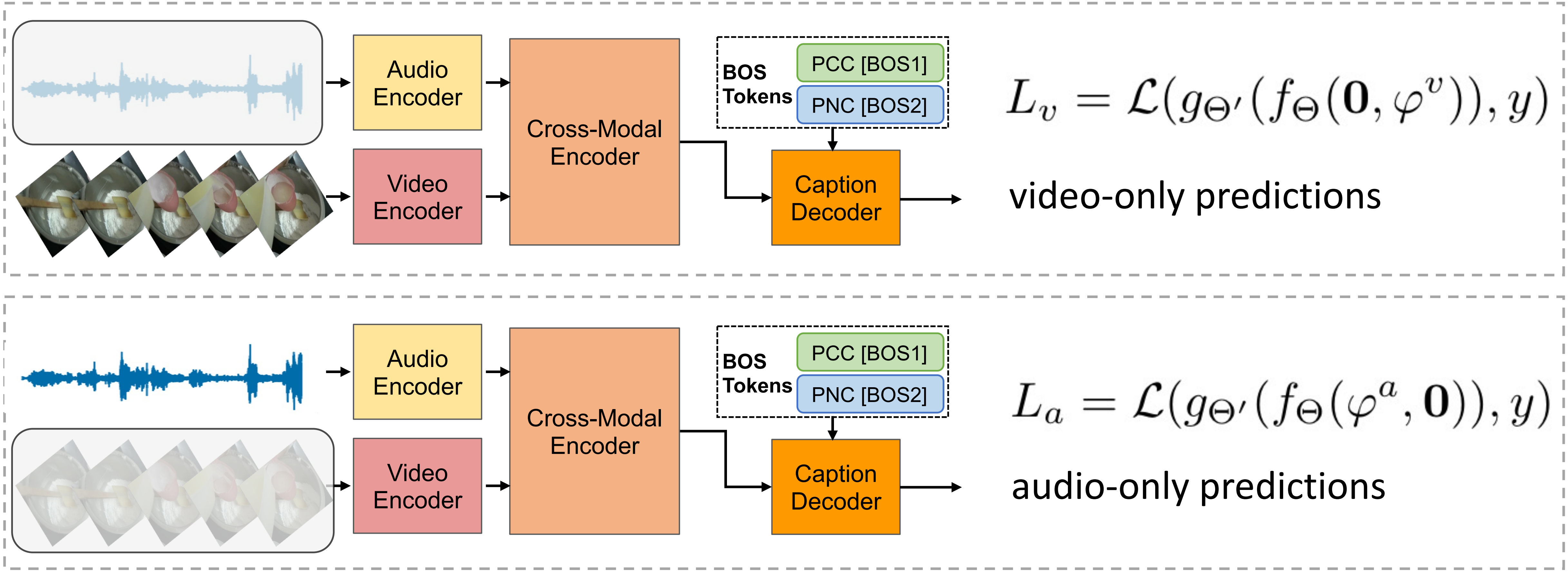
- **Audio Encoder**: Audio Spectrogram Transformer [1]
- **Video Encoder**: Video Swin Transformer [2]
- **Cross-Modal Encoder**: Local-Global Cross-Modal Fusion
- **Caption Decoder**: recursively generate captions
- **Pre-training Task**: Predict Current Caption (PCC); Predict Next Caption (PNC)



1. Y. Gong, et al. AST: Audio Spectrogram Transformer. Interspeech 2021.
2. Z. Liu, et al. Video swin transformer. CVPR 2022.

Modality Balanced Pre-training

- **Multi-modal loss:** $L = \mathcal{L}(g_{\Theta'}(f_{\Theta}(\varphi^a, \varphi^v)), y)$
- **Mono-modal losses:**



Modality Balanced Pre-training (MPB)

- **Modality Balance Pre-training**: $L_{pretrain} = L + w_a L_a + w_v L_v$
- The mono-modal weight is decided by how the modality is well utilized by model
- **Mono-to-Multi Discrepancy** (MMD) index:

$$G_a = (L_a - L)^2; G_v = (L_v - L)^2$$

- Update mono-modal weights via softmax over MMD:

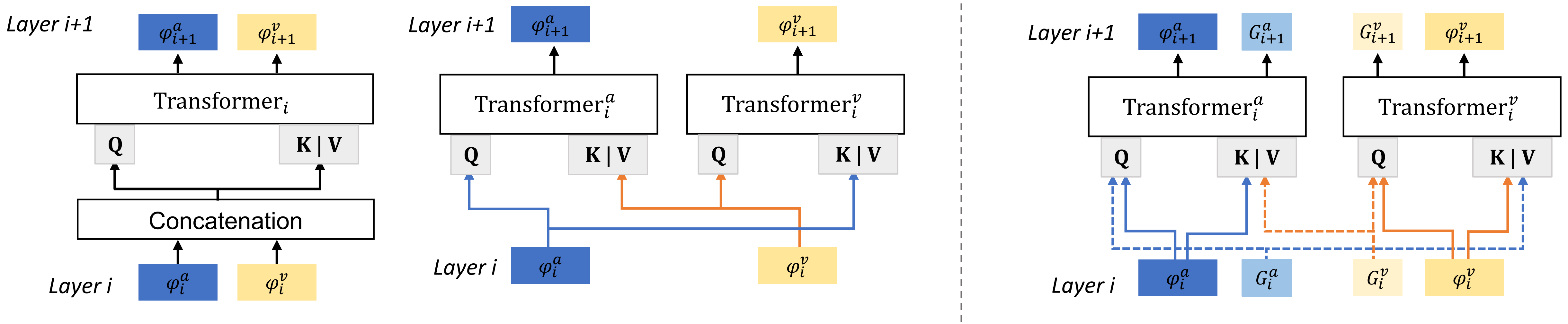
$$\tilde{w}_m^{(t)} = \frac{\exp\left(\alpha G_m^{(t)}\right)}{\sum_{m'} \exp\left(\alpha G_{m'}^{(t)}\right)}, m \in \{a, v\}$$

- Smooth during training:

$$w_m^{(t)} = \beta w_m^{(t-1)} + (1-\beta)\tilde{w}_m^{(t)}, m \in \{a, v\}$$

Cross-Modal Fusion

- **Local fusion:** merged fusion and cross fusion
  - capture local features such as words in the speech or objects in a video frame
- **Global cross fusion**: additional global tokens for cross-modal interaction
  - capture high-level concepts like sounds of laughter or people gathering on a street
- **Local-global fusion**: average of local fusion and global fusion
  - leverage multigranular information



Merged Fusion **Cross Fusion** **Global Cross Fusion**

Experiments

- MBP improves performance by a large margin on four datasets, and outperforms a multi-modal pre-training baseline G-Blend [1]
- Adding PNC leads to a remarkable boost

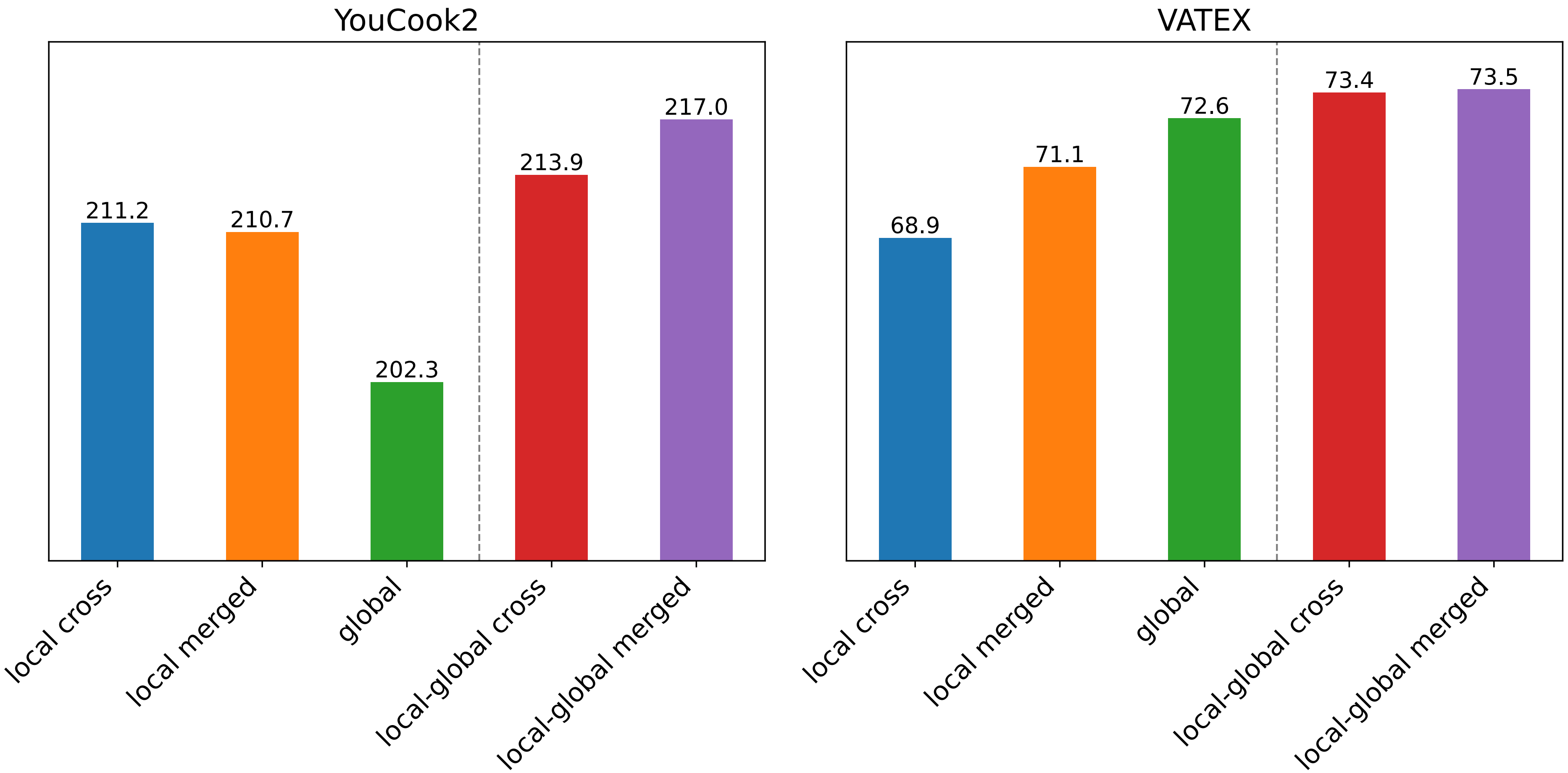
Pre-training Objective	YouCook2	MSRVTT	VATEX	ActivityNet
PCC	166.8	47.6	50.7	20.1
PCC+MBP	192.4	53.5	67.5	24.7
PCC+PNC	184.2	48.4	51.4	20.2
PCC+PNC+G-Blend [1]	208.5	55.1	68.7	25.3
PCC+PNC+MBP	**217.0**	**57.0**	**73.5**	**26.1**

Ablation studies on multi-modal pre-training. MBP: Modality Balanced Pre-training;
PCC: Predict Current Caption; PNC: Predict Next Caption.

1. W. Wang, et al. What makes training multi-modal classification networks hard? CVPR 2020.

Experiments

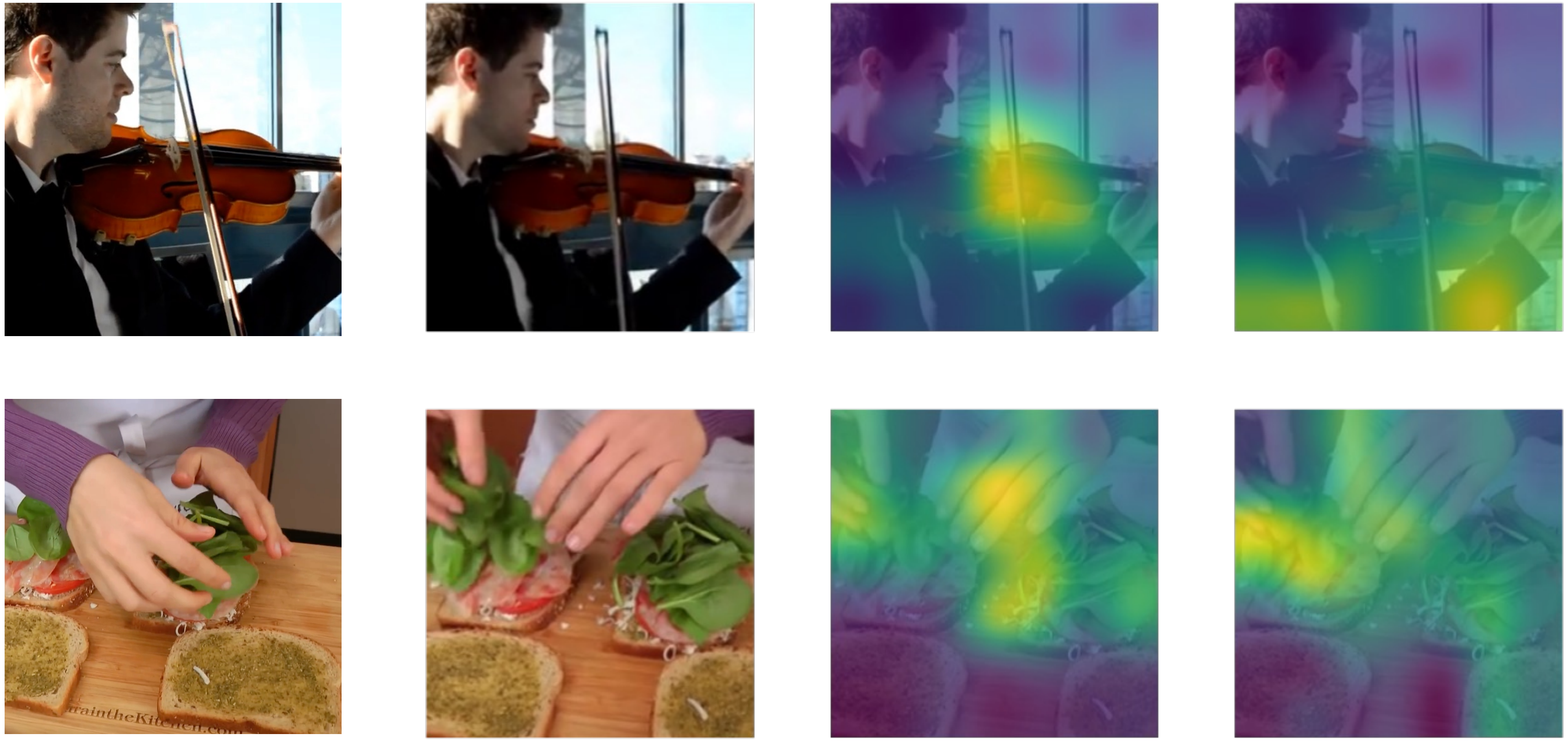
- Local-global fusion modules perform best by capturing both global and local audio information



Ablation studies on cross-modal fusion modules

Attention Maps

Video | Mid Frame | Global Fusion | Local Fusion



Caption: A man in a suit skillfully plays the violin in front of a large window.

Caption: Add spinach to the bread slices.

ASR: *“I'm just going to put on a handful of some fresh, clean baby spinach.”*

Qualitative Results

Audio Description: [Baby Crying] [Woman Speech: oh, no baby] [Woman laughter]

GroundTruth: **A baby fusses and cries** while a **woman talks** and **laughs**.
Video-only: A baby is laying down and yawning while being held by a person.
Video+Text: A baby sneezes and then sneezes several times.
Video+Audio: **A woman is laughing and talking** to a baby and **the baby is crying**.

Audio Description: [**Girl**: What's this? Pencil.] [**Man**: Pencil.] [**Girl**: What's this?] [**Man**: I don't know]

GroundTruth: **A little girl** is pointing to pictures in a book while **an adult talks to her**.
Video-only: A baby is sitting on a couch looking through a children's book.
Video+Text: A little boy is holding a pencil in front of a pencil sharpener.
Video+Audio: **A little girl** is reading a book while **a man talks to her**.

Thank you!

Exploring the Role of Audio in Video Captioning

Yuhan Shen†, Linjie Yang*, Longyin Wen*,
Haichao Yu*, Ehsan Elhamifar†, Heng Wang*
† Northeastern University * ByteDance

7th MUltimodal Learning and Applications Workshop (MULA 2024)