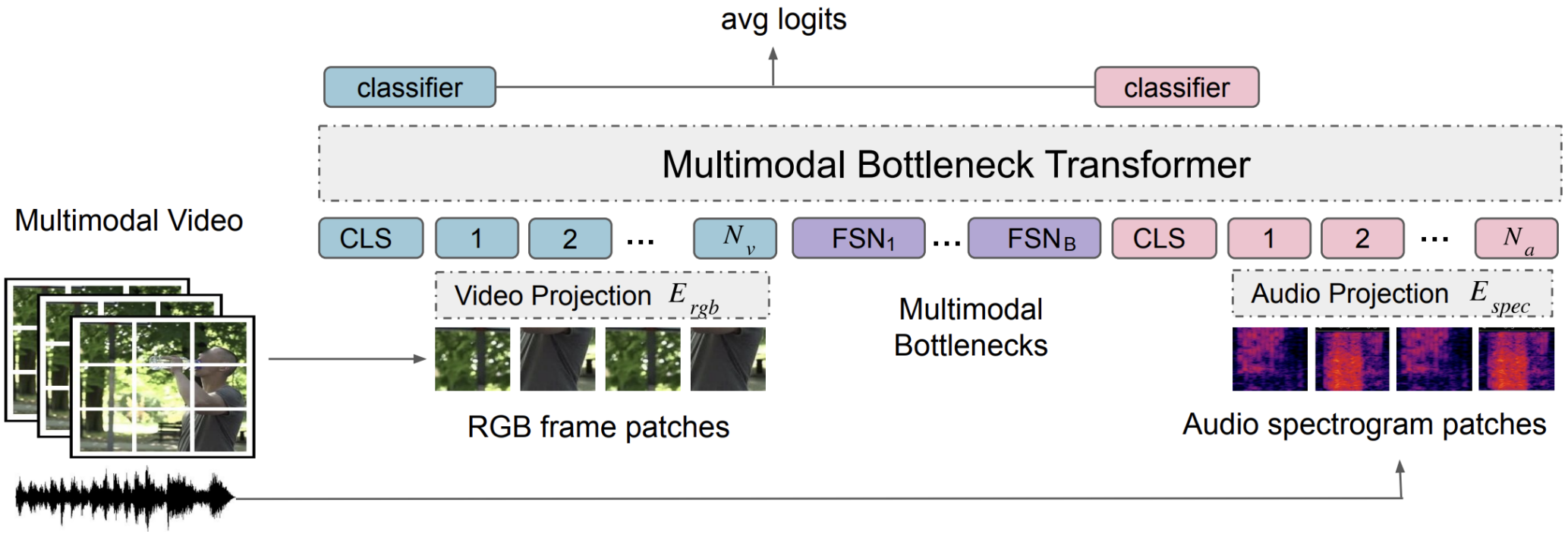
# TEST-TIME ADAPTATION FOR COMBATING MISSING MODALITIES IN EGOCENTRIC VIDEOS

**Merey Ramazanova**, Alejandro Pardo, Bernard Ghanem, Motasem Alfarra

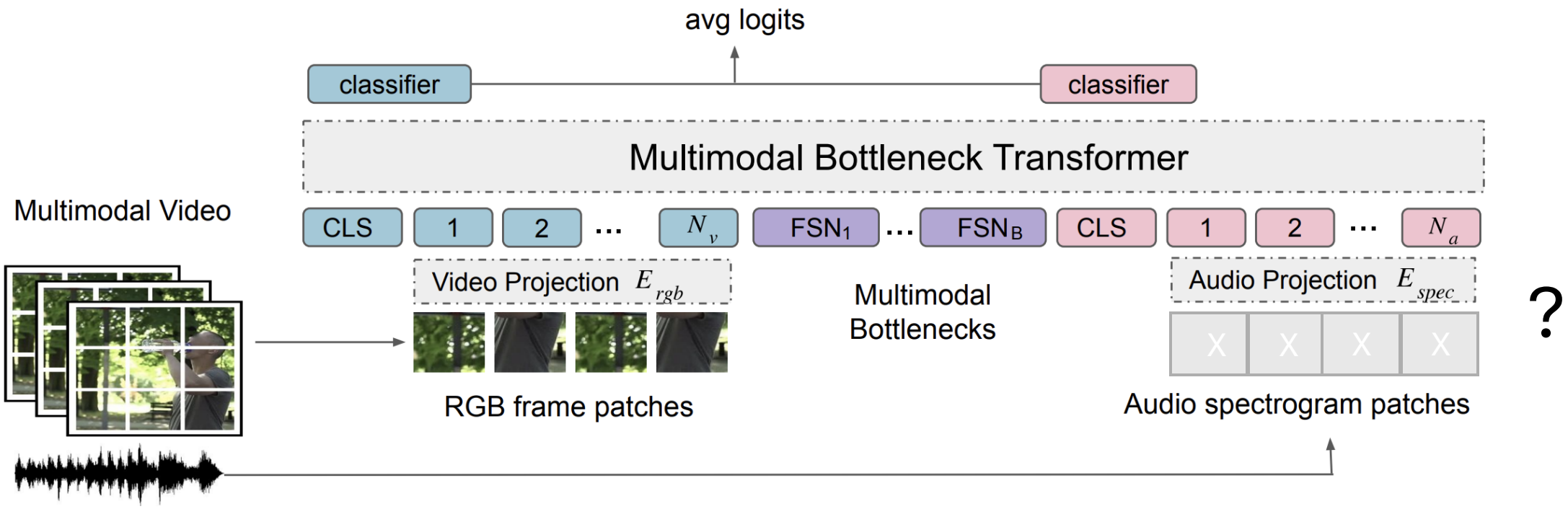
Center of Excellence in Generative AI, KAUST, Saudi Arabia

# Challenges of missing modalities in multimodal learning




Nagrani, Arsha, et al. "Attention bottlenecks for multimodal fusion." *Advances in neural information processing systems* 34 (2021): 14200-14213.

# Challenges of missing modalities in multimodal learning




Nagrani, Arsha, et al. "Attention bottlenecks for multimodal fusion." *Advances in neural information processing systems* 34 (2021): 14200-14213.

# Challenges of missing modalities in multimodal learning

# Why missing modalities

- Wearable devices:

# Why missing modalities

- Wearable devices:
  - Privacy

# Why missing modalities

- Wearable devices:
  - Privacy
  - Cost

# Why missing modalities

- Wearable devices:
  - Privacy
  - Cost

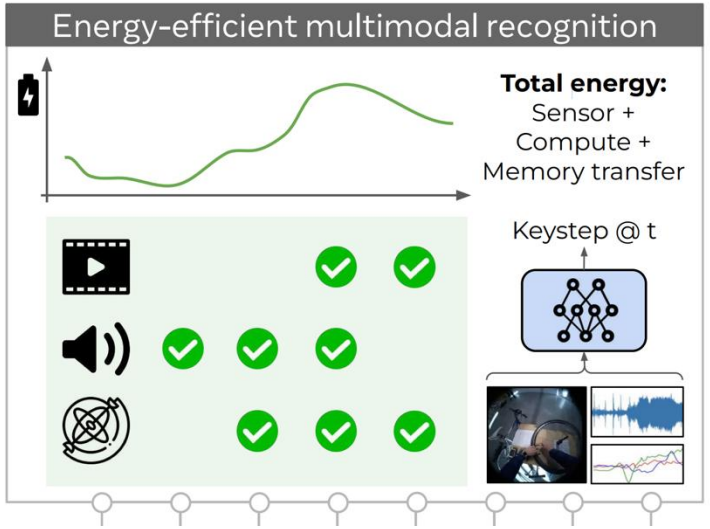
| Modality: | RGB video | Text narrations | Features | Audio | Faces | 3D scans | Stereo | Gaze | IMU | Multi-cam |
|---|---|---|---|---|---|---|---|---|---|---|
| # hours: | 3,670 | 3,670 | 3,670 | 2,535 | 612 | 491 | 80 | 45 | 836 | 224 |

Table 1. Modalities of data in Ego4D and their amounts. "Narrations" are dense, timestamped descriptions of camera wearer activity (cf. Sec. 4). "3D scans" are meshes from Matterport3D scanners for the full environment in which the video was captured. "Faces" refers to video where participants consented to remain unblurred. "Multi-cam" refers to synchronized video captured at the same event by multiple camera wearers. "Features" refers to precomputed SlowFast [70] video features. Gaze collected only by Indiana U. and Georgia Tech.

Grauman, Kristen, et al. "Ego4d: Around the world in 3,000 hours of egocentric video." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

# Why missing modalities

- Wearable devices:
  - Privacy
  - Cost




| Modality: | RGB video | Text narrations | Features | Audio | Faces | 3D scans | Stereo | Gaze | IMU | Multi-cam |
|---|---|---|---|---|---|---|---|---|---|---|
| # hours: | 3,670 | 3,670 | 3,670 | 2,535 | 612 | 491 | 80 | 45 | 836 | 224 |

Table 1. Modalities of data in Ego4D and their amounts. "Narrations" are dense, timestamped descriptions of camera wearer activity (cf. Sec. 4). "3D scans" are meshes from Matterport3D scanners for the full environment in which the video was captured. "Faces" refers to video where participants consented to remain unblurred. "Multi-cam" refers to synchronized video captured at the same event by multiple camera wearers. "Features" refers to precomputed SlowFast [70] video features. Gaze collected only by Indiana U. and Georgia Tech.

Grauman, Kristen, et al. "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

# Prior approaches to handle missing modalities

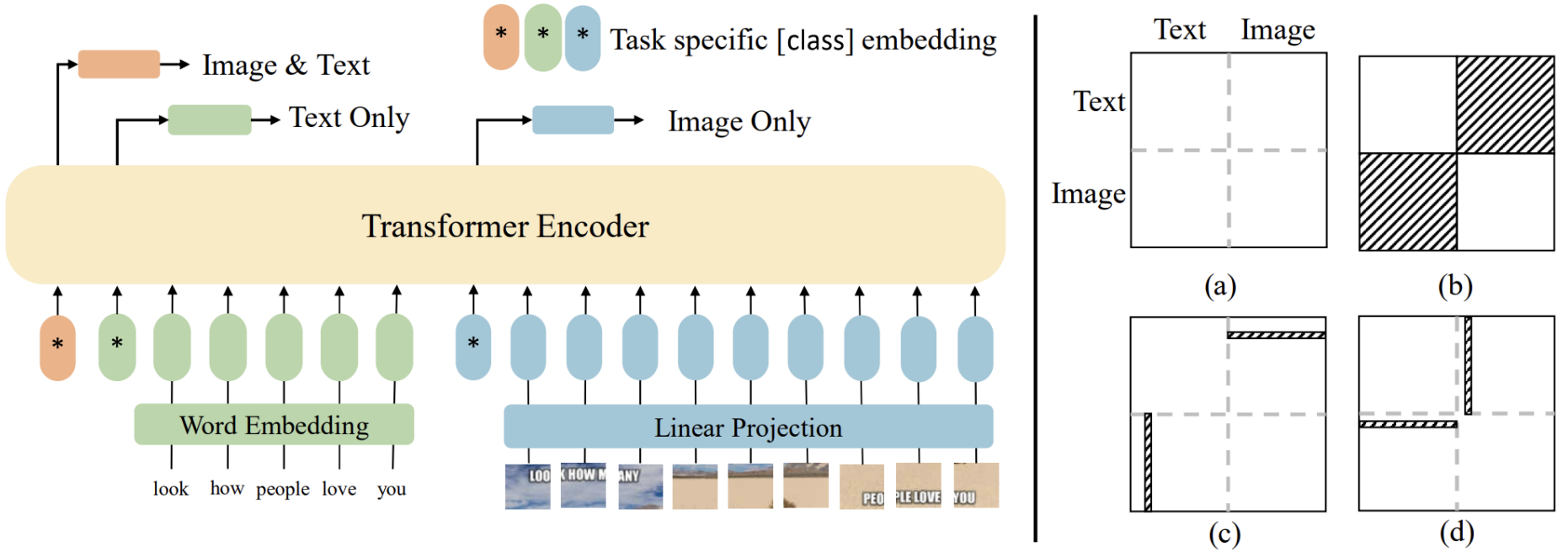



Figure 1. *Left*: Overview of our model. *Right*: Attention masks for different tasks: (a) Original attention without masking; (b) Mask-out cross-modal attention; (c) Mask-out image attention for text only [class] token; (d) Mask-out text attention for image only [class] token.

Ma, Mengmeng, et al. "Are multimodal transformers robust to missing modality?." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

# Prior approaches to handle missing modalities

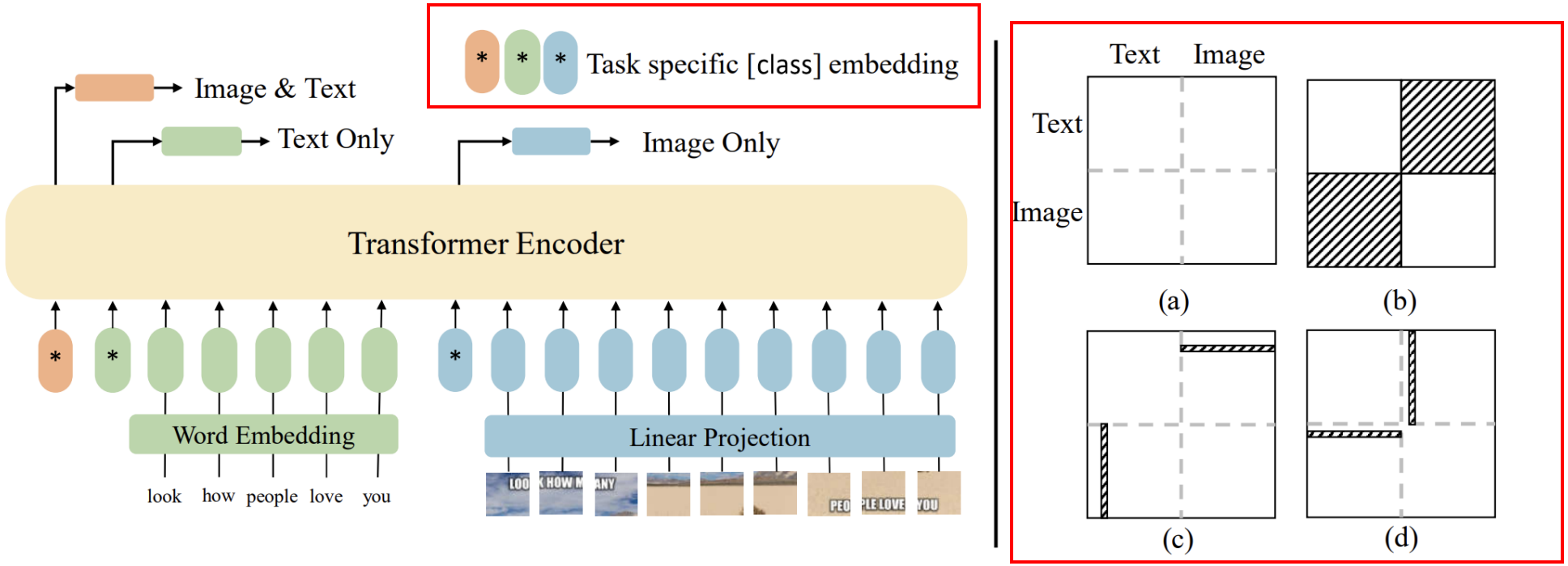



Figure 1. *Left*: Overview of our model. *Right*: Attention masks for different tasks: (a) Original attention without masking; (b) Mask-out cross-modal attention; (c) Mask-out image attention for text only [class] token; (d) Mask-out text attention for image only [class] token.

Ma, Mengmeng, et al. "Are multimodal transformers robust to missing modality?." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

# Prior approaches to handle missing modalities

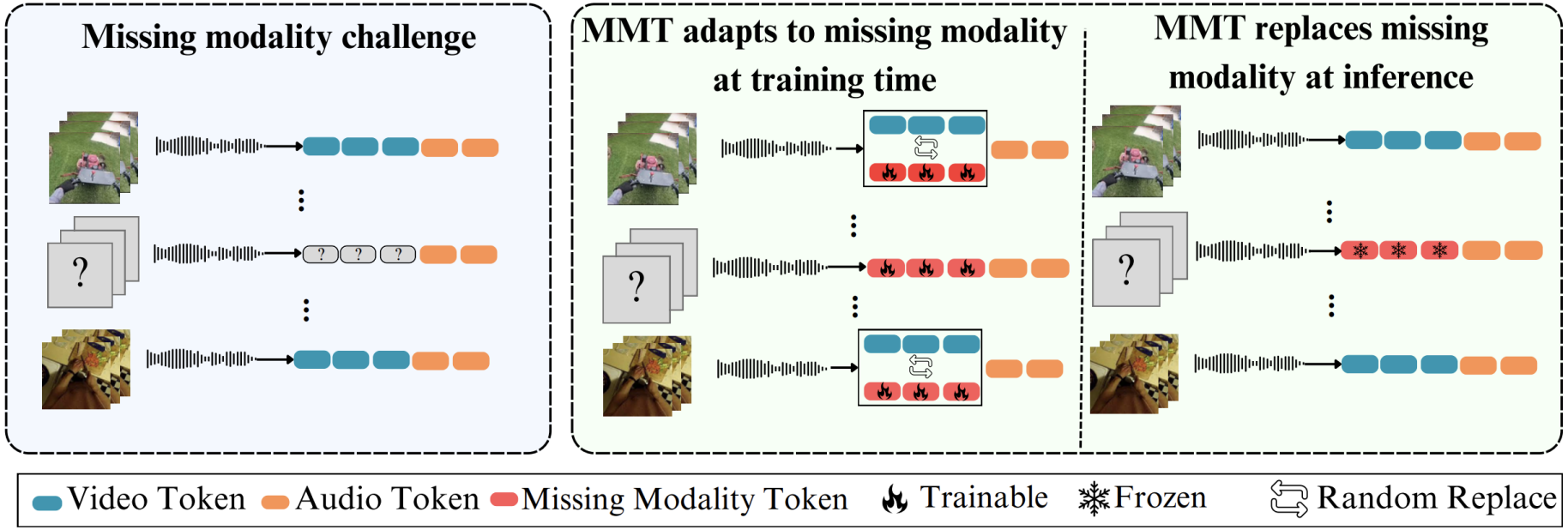



Figure 2. **Learning and Predicting with Missing Modalities. Left:** Given modal-incomplete data, it is still unclear how to effectively train and predict with a multimodal model (we present some naive baseline methods in Sec. 3.3). **Right:** To address this issue, we introduce a Missing Modality Token (MMT). During training, MMT learns the representation of missing inputs from modal-incomplete samples and modal-complete samples. For the latter, we use *random-replace* to let the network observe the missing inputs and thus learn better representations (Sec. 3.4). At test time, we replace the tokens of missing inputs with MMT to effectively represent them.

Ramazanova, Merey, et al. "Exploring missing modality in multimodal egocentric datasets." *arXiv preprint arXiv:2401.11470* (2024).

# Prior approaches to handle missing modalities

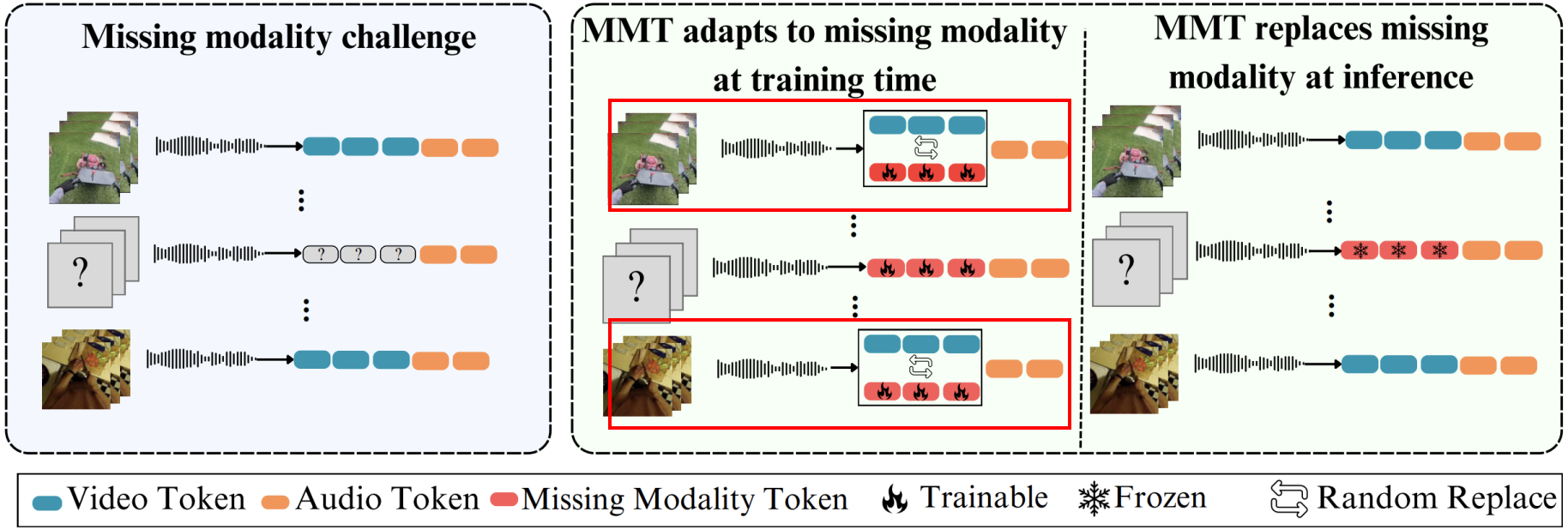



Figure 2. **Learning and Predicting with Missing Modalities. Left:** Given modal-incomplete data, it is still unclear how to effectively train and predict with a multimodal model (we present some naive baseline methods in Sec. 3.3). **Right:** To address this issue, we introduce a Missing Modality Token (MMT). During training, MMT learns the representation of missing inputs from modal-incomplete samples and modal-complete samples. For the latter, we use *random-replace* to let the network observe the missing inputs and thus learn better representations (Sec. 3.4). At test time, we replace the tokens of missing inputs with MMT to effectively represent them.

Ramazanova, Merey, et al. "Exploring missing modality in multimodal egocentric datasets." *arXiv preprint arXiv:2401.11470* (2024).

# Prior approaches to handle missing modalities

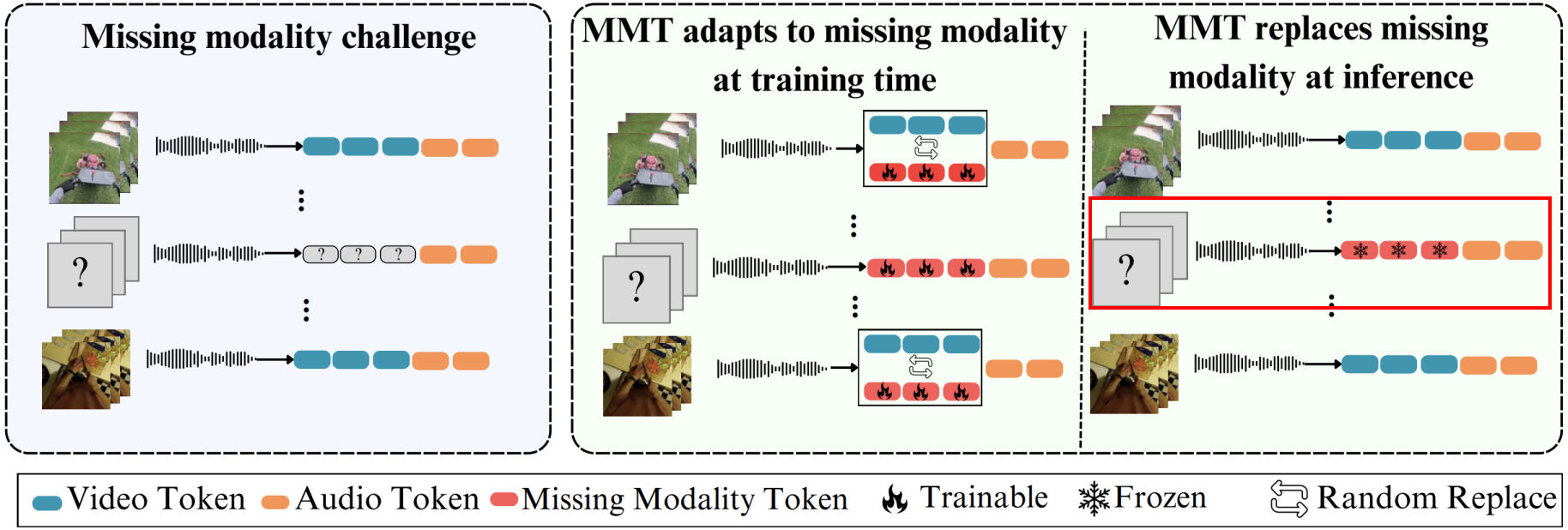



Figure 2. **Learning and Predicting with Missing Modalities. Left:** Given modal-incomplete data, it is still unclear how to effectively train and predict with a multimodal model (we present some naive baseline methods in Sec. 3.3). **Right:** To address this issue, we introduce a Missing Modality Token (MMT). During training, MMT learns the representation of missing inputs from modal-incomplete samples and modal-complete samples. For the latter, we use *random-replace* to let the network observe the missing inputs and thus learn better representations (Sec. 3.4). At test time, we replace the tokens of missing inputs with MMT to effectively represent them.

Ramazanova, Merey, et al. "Exploring missing modality in multimodal egocentric datasets." *arXiv preprint arXiv:2401.11470* (2024).

# Prior approaches to handle missing modalities

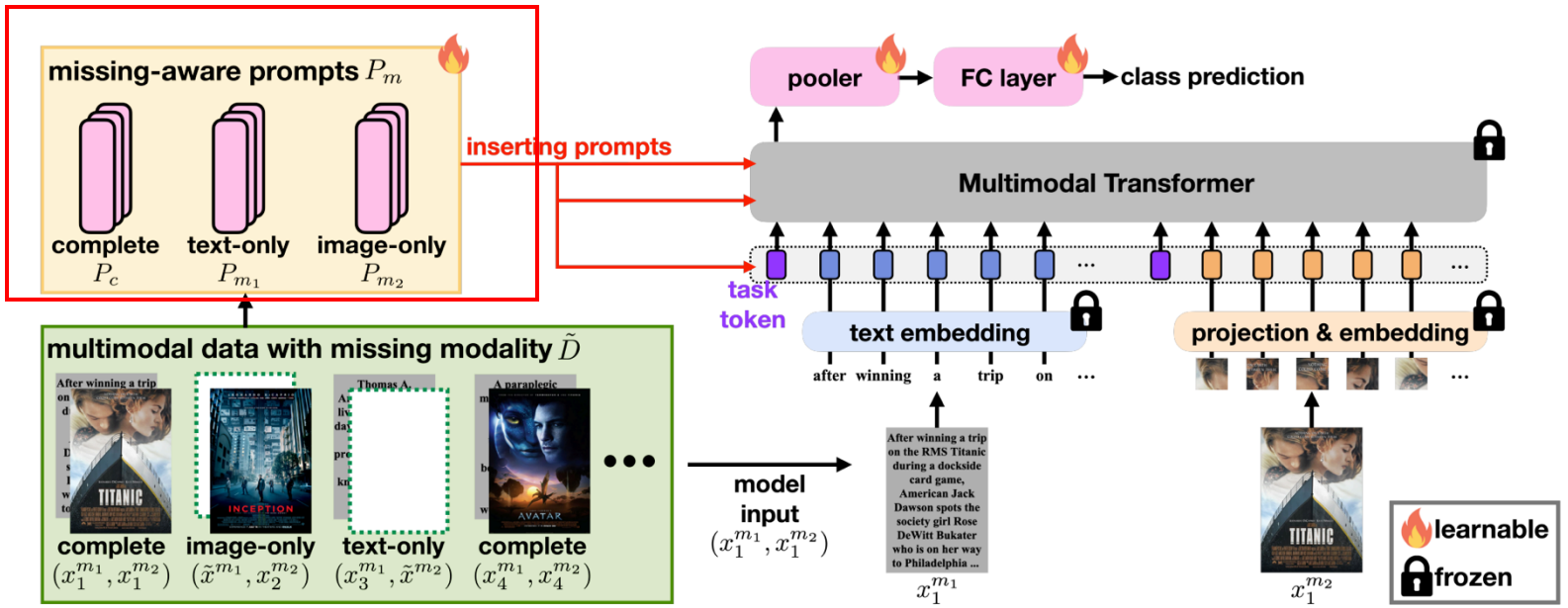



Figure 2. The overview of our proposed prompt-based multimodal framework. We first select the missing-aware prompts $P_m$ according to the missing case (e.g., complete, text-only, image-only in vision-language tasks) of the multimodal inputs $(x_i^{m_1}, x_i^{m_2})$, in which the dummy inputs $\{\tilde{x}^{m_1}, \tilde{x}^{m_2}\}$ respectively for text and image are adopted for the corresponding missing modality. Then we attach missing-aware prompts into multiple MSA layers via different prompting approaches (see Figure 3 and Section 3.3). We select the text-related task token of the multimodal transformer as our final output features, and feed them to the pooler layer and fully-connected (FC) layers for class predictions. Note that only the pink-shaded blocks require to be trained while the others are frozen.

Lee, Yi-Lun, et al. "Multimodal prompting with missing modalities for visual recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

# Missing modality as a test-time adaptation (TTA) challenge

*Can we develop methods to address missing modalities at test time without imposing retraining requirements?*

# Missing modality as a test-time adaptation (TTA) challenge

*Can we develop methods to address missing modalities at test time without imposing retraining requirements?*

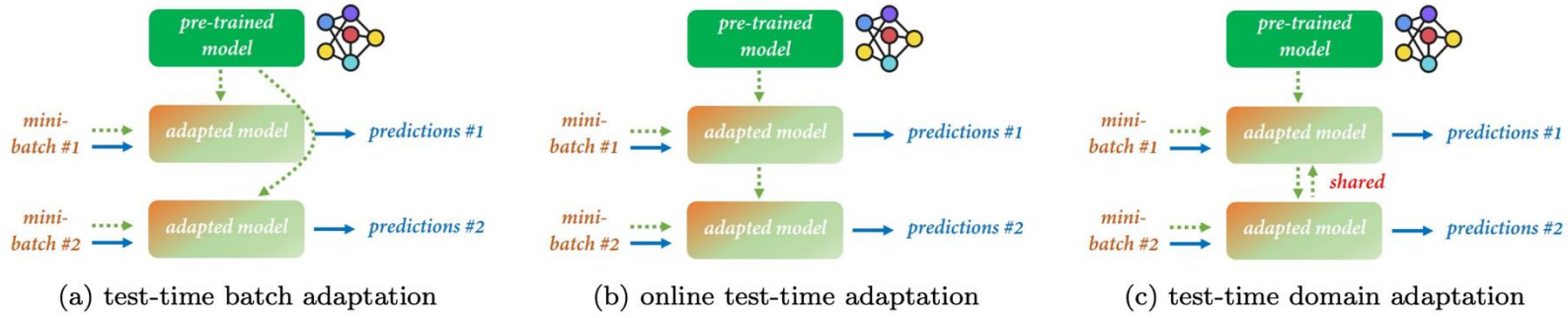



**Fig. 1** The **test-time adaptation (TTA) paradigm** aims to adapt the pre-trained model to various types of unlabeled test data, including single mini-batch in (a), streaming data in (b), or an entire dataset in (c), before making predictions. During the adaptation process, either the model or the input data can be altered to improve performance against distribution shifts. The dotted green arrow indicates the test-time training phase before inference, while the blue arrow denotes pure inference.

Liang, Jian, Ran He, and Tieniu Tan. "A comprehensive survey on test-time adaptation under distribution shifts." International Journal of Computer Vision 133.1 (2025): 31-64.

# Missing modality as a test-time adaptation (TTA) challenge

*Can we develop methods to address missing modalities at test time without imposing retraining requirements?*

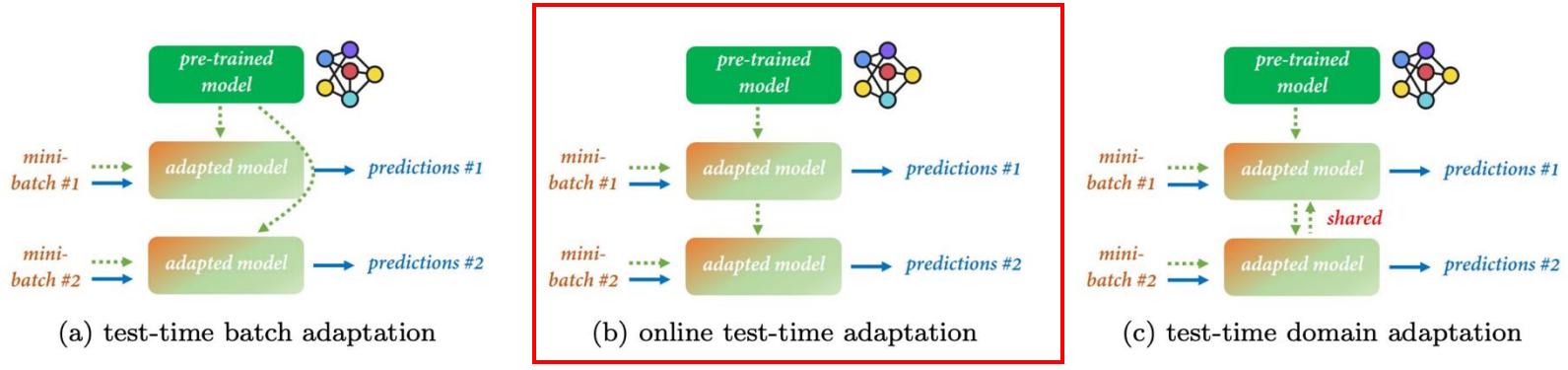



**Fig. 1** The **test-time adaptation (TTA) paradigm** aims to adapt the pre-trained model to various types of unlabeled test data, including single mini-batch in (a), streaming data in (b), or an entire dataset in (c), before making predictions. During the adaptation process, either the model or the input data can be altered to improve performance against distribution shifts. The dotted green arrow indicates the test-time training phase before inference, while the blue arrow denotes pure inference.

Liang, Jian, Ran He, and Tieniu Tan. "A comprehensive survey on test-time adaptation under distribution shifts." International Journal of Computer Vision 133.1 (2025): 31-64.

# Missing modality as a test-time adaptation (TTA) challenge


Figure 1: **Test-Time Adaptation for missing modalities.** The concept of test-time adaptation in the presence of missing data modalities focuses on a system where a stream of multimodal data is input, potentially lacking one or more modalities. Without adaptation, the pretrained model $f_{\theta_0}$ may predict inaccurate labels due to incomplete data. With test-time adaptation, the model is dynamically adjusted using the adaptation method $g$, resulting in an adapted model $f_{\theta_t}$, designed to handle the missing modalities and improve over time. The graph on the right illustrates the performance of the non-adapted baseline (blue) vs. the model adapted with our proposed adaptation method MiDl (green) on Epic-Kitchens dataset. It shows the adaptation efficacy in maintaining higher performance levels despite the variability in modal-completeness, surpassing the unimodal performance (orange) for all missing rates.

# Missing modality as a test-time adaptation (TTA) challenge


Figure 1: **Test-Time Adaptation for missing modalities.** The concept of test-time adaptation in the presence of missing data modalities focuses on a system where a stream of multimodal data is input, potentially lacking one or more modalities. Without adaptation, the pretrained model $f_{\theta_0}$ may predict inaccurate labels due to incomplete data. With test-time adaptation, the model is dynamically adjusted using the adaptation method $g$, resulting in an adapted model $f_{\theta_t}$, designed to handle the

1. $\mathcal{S}$ reveals a sample/batch $x_t$ with its corresponding modality $m$.
2. $f_{\theta_t}$ generates the prediction $\hat{y}_t$.
3. $g$ adapts the model parameter $\theta_t$ to $\theta_{t+1}$.

# Missing modality as a test-time adaptation (TTA) challenge

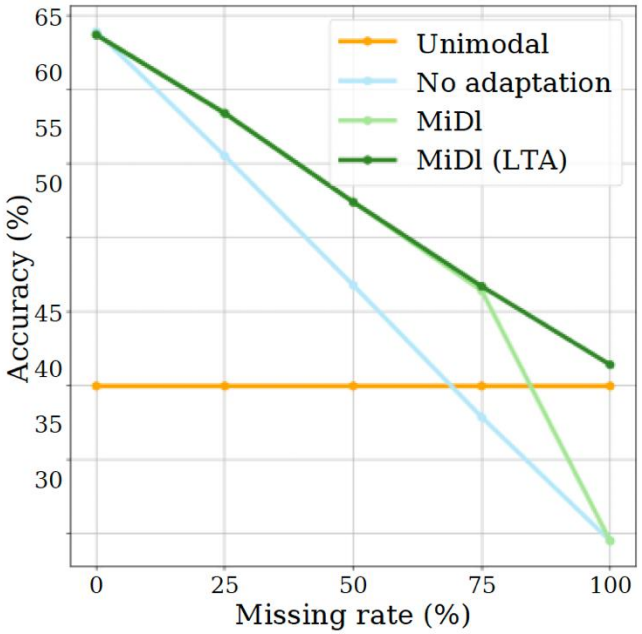




Figure 1: **Test-Time Adaptation for missing modalities.** Th
the presence of missing data modalities focuses on a system v
input, potentially lacking one or more modalities. Without adap
predict inaccurate labels due to incomplete data. With test-time
adjusted using the adaptation method $g$, resulting in an adapte
missing modalities and improve over time. The graph on the
the non-adapted baseline (blue) vs. the model adapted with ou
(green) on Epic-Kitchens dataset. It shows the adaptation efficacy in maintaining higher performance levels despite the variability in modal-completeness, surpassing the unimodal performance (orange) for all missing rates.

# Our solution: MiDl

- How should an optimal $f_\theta$ behave under missing modality?

# Our solution: MiDl

- How should an optimal $f_\theta$ behave under missing modality?
  - the prediction of $f_\theta$ should be invariant to the modality source *m*. $f_\theta$ *should output the same prediction under both complete and incomplete modality, hence satisfying the following equality:*

$$f_\theta(x_i; M = A) = f_\theta(x_i; M = V) = f_\theta(x_i; M = AV)\,\forall i.$$

# Our solution: MiDl

- How should an optimal $f_\theta$ behave under missing modality?
  - the prediction of $f_\theta$ should be invariant to the modality source *m*. $f_\theta$ *should output the same prediction under both complete and incomplete modality, hence satisfying the following equality:*

$$f_\theta(x_i; M = A) = f_\theta(x_i; M = V) = f_\theta(x_i; M = AV)\,\forall i.$$

$$\theta^* = \arg\min_\theta \mathbb{E}_{x\sim\mathcal{S}} [\mathbf{MI}\left(f_\theta(x; m), m\right) + \mathbf{KL}\left(f_\theta(x \mid M = AV)||f_{\theta_0}(x \mid M = AV)\right]$$

# Our solution: MiDl

- How should an optimal $f_\theta$ behave under missing modality?
  - the prediction of $f_\theta$ should be invariant to the modality source *m*. $f_\theta$ *should output the same prediction under both complete and incomplete modality, hence satisfying the following equality:*

$$f_\theta(x_i; M = A) = f_\theta(x_i; M = V) = f_\theta(x_i; M = AV)\,\forall i.$$

  - $f_\theta$ *should retain high performance in predicting data with complete modality, which is generally satisfied for* $f_\theta$

$$\theta^* = \arg\min_\theta \mathbb{E}_{x\sim\mathcal{S}} \left[\mathbf{MI}\left(f_\theta(x; m), m\right) + \mathbf{KL}\left(f_\theta(x \mid M = AV) || f_{\theta_0}(x \mid M = AV)\right)\right]$$

# Our solution: MiDl

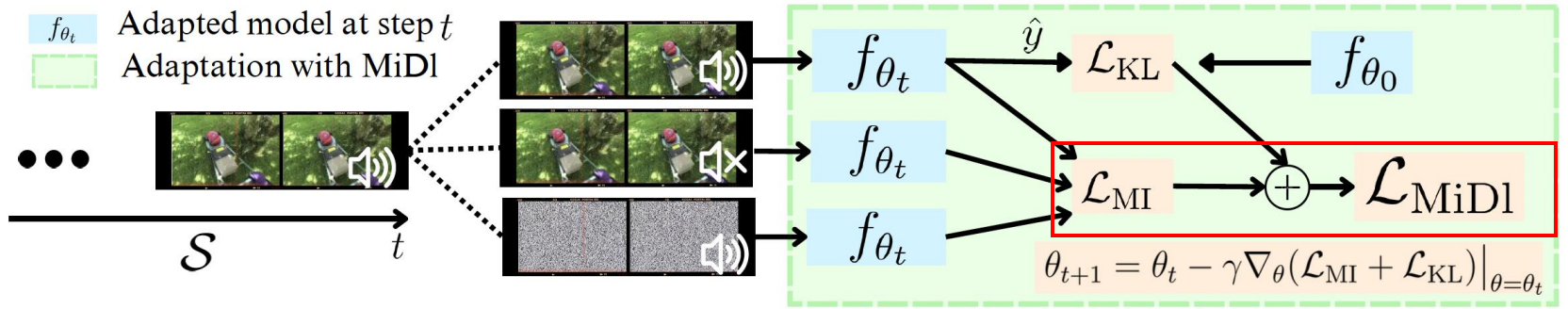



Figure 2: **Adapting at test-time with MiDl.** At test time, the stream reveals a sample. MiDl uses multimodal samples to adapt and requires one forward pass for each modality combination. MiDl leverages (KL) divergence to align the predictions of the adapted model $f_{\theta_t}$ with those of the original model $f_{\theta_0}$, ensuring that the adapted model does not deviate too far from the original model's predictions. The Mutual-Information (MI) component uses the prediction from the different modalities to reduce the dependency on any specific modality, fostering a more generalized and robust prediction across different modality combinations. MiDl updates the model for step $t + 1$ using the combination of KL and MI in Equation 2.

the prediction of $f_\theta$ should be invariant to the modality source

# Our solution: MiDl

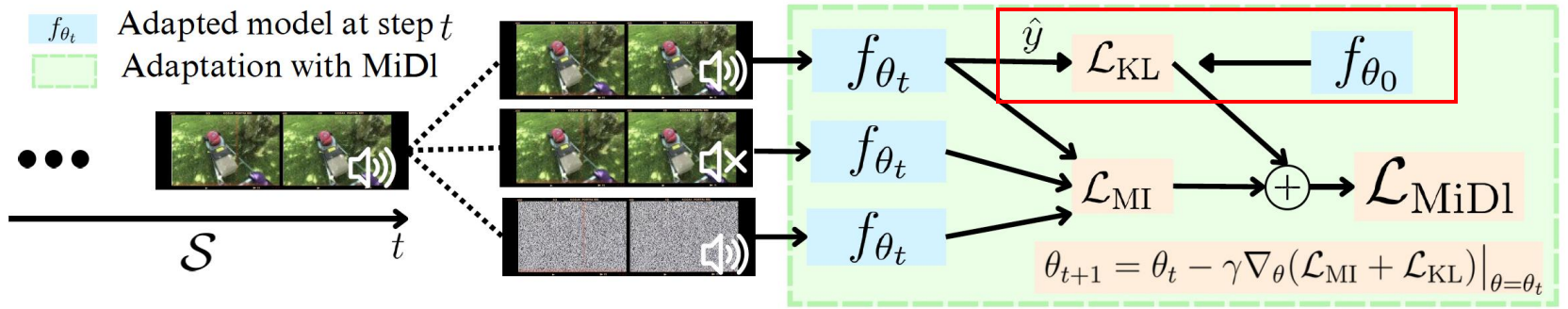



Figure 2: **Adapting at test-time with MiDl.** At test time, the stream reveals a sample. MiDl uses multimodal samples to adapt and requires one forward pass for each modality combination. MiDl leverages (KL) divergence to align the predictions of the adapted model $f_{\theta_t}$ with those of the original model $f_{\theta_0}$, ensuring that the adapted model does not deviate too far from the original model's predictions. The Mutual-Information (MI) component uses the prediction from the different modalities to reduce the dependency on any specific modality, fostering a more generalized and robust prediction across different modality combinations. MiDl updates the model for step $t + 1$ using the combination of KL and MI in Equation 2.

$f_\theta$ *should retain high performance in predicting data with complete modality*

# Results: MiDl

Table 1: **Combating missing modalities at test time.** The first two rows show the unimodal performance and the MBT baseline with no adaptation. We show three alternative TTA methods and demonstrate that our proposed MiDl is effective at combating missing modalities at test time, outperforming all presented TTA baselines. Refer to Table 11 to see the standard deviations.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) | | | | | Epic-Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 41.4 | 41.4 | 41.4 | 41.4 | 41.4 | 40.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| Baseline | 55.1 | 45.6 | 37.1 | 28.3 | 19.5 | 63.9 | 55.5 | 46.8 | 37.9 | 29.5 |
| +Shot | 55.0 | 45.6 | 37.1 | 28.5 | **20.0** | **63.9** | 55.9 | 47.9 | 40.6 | **34.3** |
| +Tent | 54.8 | 45.0 | 35.9 | 26.5 | 17.8 | 63.7 | 54.0 | 39.2 | 24.2 | 9.9 |
| +ETA | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | 63.5 | 51.3 | 33.7 | 20.6 | 7.9 |
| +MiDl (ours) | 55.0 | **46.8** | **38.8** | **29.8** | 19.5 | 63.7 | **58.4** | **52.4** | **46.4** | 29.5 |

# Results: MiDl

Table 1: **Combating missing modalities at test time.** The first two rows show the unimodal performance and the MBT baseline with no adaptation. We show three alternative TTA methods and demonstrate that our proposed MiDl is effective at combating missing modalities at test time, outperforming all presented TTA baselines. Refer to Table 11 to see the standard deviations.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) | | | | | Epic-Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 41.4 | 41.4 | 41.4 | 41.4 | 41.4 | 40.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| Baseline | 55.1 | 45.6 | 37.1 | 28.3 | 19.5 | 63.9 | 55.5 | 46.8 | 37.9 | 29.5 |
| +Shot | 55.0 | 45.6 | 37.1 | 28.5 | **20.0** | **63.9** | 55.9 | 47.9 | 40.6 | **34.3** |
| +Tent | 54.8 | 45.0 | 35.9 | 26.5 | 17.8 | 63.7 | 54.0 | 39.2 | 24.2 | 9.9 |
| +ETA | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | 63.5 | 51.3 | 33.7 | 20.6 | 7.9 |
| +MiDl (ours) | 55.0 | **46.8** | **38.8** | **29.8** | 19.5 | 63.7 | **58.4** | **52.4** | **46.4** | 29.5 |

# Results: MiDl

Table 1: **Combating missing modalities at test time.** The first two rows show the unimodal performance and the MBT baseline with no adaptation. We show three alternative TTA methods and demonstrate that our proposed MiDl is effective at combating missing modalities at test time, outperforming all presented TTA baselines. Refer to Table 11 to see the standard deviations.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) | | | | | Epic-Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 41.4 | 41.4 | 41.4 | 41.4 | 41.4 | 40.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| Baseline | 55.1 | 45.6 | 37.1 | 28.3 | 19.5 | 63.9 | 55.5 | 46.8 | 37.9 | 29.5 |
| +Shot | 55.0 | 45.6 | 37.1 | 28.5 | **20.0** | **63.9** | 55.9 | 47.9 | 40.6 | **34.3** |
| +Tent | 54.8 | 45.0 | 35.9 | 26.5 | 17.8 | 63.7 | 54.0 | 39.2 | 24.2 | 9.9 |
| +ETA | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | 63.5 | 51.3 | 33.7 | 20.6 | 7.9 |
| +MiDl (ours) | 55.0 | **46.8** | **38.8** | **29.8** | 19.5 | 63.7 | **58.4** | **52.4** | **46.4** | 29.5 |

# Results: MiDl

Table 1: **Combating missing modalities at test time.** The first two rows show the unimodal performance and the MBT baseline with no adaptation. We show three alternative TTA methods and demonstrate that our proposed MiDl is effective at combating missing modalities at test time, outperforming all presented TTA baselines. Refer to Table 11 to see the standard deviations.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) | | | | | Epic-Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 41.4 | 41.4 | 41.4 | 41.4 | 41.4 | 40.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| Baseline | 55.1 | 45.6 | 37.1 | 28.3 | 19.5 | 63.9 | 55.5 | 46.8 | 37.9 | 29.5 |
| +Shot | 55.0 | 45.6 | 37.1 | 28.5 | **20.0** | **63.9** | 55.9 | 47.9 | 40.6 | **34.3** |
| +Tent | 54.8 | 45.0 | 35.9 | 26.5 | 17.8 | 63.7 | 54.0 | 39.2 | 24.2 | 9.9 |
| +ETA | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | 63.5 | 51.3 | 33.7 | 20.6 | 7.9 |
| +MiDl (ours) | 55.0 | **46.8** | **38.8** | **29.8** | 19.5 | 63.7 | **58.4** | **52.4** | **46.4** | 29.5 |

# Results: MiDl

Table 1: **Combating missing modalities at test time.** The first two rows show the unimodal performance and the MBT baseline with no adaptation. We show three alternative TTA methods and demonstrate that our proposed MiDl is effective at combating missing modalities at test time, outperforming all presented TTA baselines. Refer to Table 11 to see the standard deviations.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) | | | | | Epic-Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 41.4 | 41.4 | 41.4 | 41.4 | 41.4 | 40.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| Baseline | 55.1 | 45.6 | 37.1 | 28.3 | 19.5 | 63.9 | 55.5 | 46.8 | 37.9 | 29.5 |
| +Shot | 55.0 | 45.6 | 37.1 | 28.5 | **20.0** | **63.9** | 55.9 | 47.9 | 40.6 | **34.3** |
| +Tent | 54.8 | 45.0 | 35.9 | 26.5 | 17.8 | 63.7 | 54.0 | 39.2 | 24.2 | 9.9 |
| +ETA | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | 63.5 | 51.3 | 33.7 | 20.6 | 7.9 |
| +MiDl (ours) | 55.0 | **46.8** | **38.8** | **29.8** | 19.5 | 63.7 | **58.4** | **52.4** | **46.4** | 29.5 |

# Results: MiDl

Table 1: **Combating missing modalities at test time.** The first two rows show the unimodal performance and the MBT baseline with no adaptation. We show three alternative TTA methods and demonstrate that our proposed MiDl is effective at combating missing modalities at test time, outperforming all presented TTA baselines. Refer to Table 11 to see the standard deviations.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) | | | | | Epic-Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 41.4 | 41.4 | 41.4 | 41.4 | 41.4 | 40.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| Baseline | 55.1 | 45.6 | 37.1 | 28.3 | 19.5 | 63.9 | 55.5 | 46.8 | 37.9 | 29.5 |
| +Shot | 55.0 | 45.6 | 37.1 | 28.5 | **20.0** | **63.9** | 55.9 | 47.9 | 40.6 | **34.3** |
| +Tent | 54.8 | 45.0 | 35.9 | 26.5 | 17.8 | 63.7 | 54.0 | 39.2 | 24.2 | 9.9 |
| +ETA | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | 63.5 | 51.3 | 33.7 | 20.6 | 7.9 |
| +MiDl (ours) | 55.0 | **46.8** | **38.8** | **29.8** | 19.5 | 63.7 | **58.4** | **52.4** | **46.4** | 29.5 |

# Results: MiDl

Table 1: **Combating missing modalities at test time.** The first two rows show the unimodal performance and the MBT baseline with no adaptation. We show three alternative TTA methods and demonstrate that our proposed MiDl is effective at combating missing modalities at test time, outperforming all presented TTA baselines. Refer to Table 11 to see the standard deviations.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) | | | | | Epic-Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 41.4 | 41.4 | 41.4 | 41.4 | 41.4 | 40.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| Baseline | 55.1 | 45.6 | 37.1 | 28.3 | 19.5 | 63.9 | 55.5 | 46.8 | 37.9 | 29.5 |
| +Shot | 55.0 | 45.6 | 37.1 | 28.5 | **20.0** | **63.9** | 55.9 | 47.9 | 40.6 | **34.3** |
| +Tent | 54.8 | 45.0 | 35.9 | 26.5 | 17.8 | 63.7 | 54.0 | 39.2 | 24.2 | 9.9 |
| +ETA | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | 63.5 | 51.3 | 33.7 | 20.6 | 7.9 |
| +MiDl (ours) | 55.0 | **46.8** | **38.8** | **29.8** | 19.5 | 63.7 | **58.4** | **52.4** | **46.4** | 29.5 |

# Results: MiDl

Table 1: **Combating missing modalities at test time.** The first two rows show the unimodal performance and the MBT baseline with no adaptation. We show three alternative TTA methods and demonstrate that our proposed MiDl is effective at combating missing modalities at test time, outperforming all presented TTA baselines. Refer to Table 11 to see the standard deviations.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) | | | | | Epic-Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 41.4 | 41.4 | 41.4 | 41.4 | 41.4 | 40.0 | 40.0 | 40.0 | 40.0 | 40.0 |
| Baseline | 55.1 | 45.6 | 37.1 | 28.3 | 19.5 | 63.9 | 55.5 | 46.8 | 37.9 | 29.5 |
| +Shot | 55.0 | 45.6 | 37.1 | 28.5 | **20.0** | **63.9** | 55.9 | 47.9 | 40.6 | **34.3** |
| +Tent | 54.8 | 45.0 | 35.9 | 26.5 | 17.8 | 63.7 | 54.0 | 39.2 | 24.2 | 9.9 |
| +ETA | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | 63.5 | 51.3 | 33.7 | 20.6 | 7.9 |
| +MiDl (ours) | 55.0 | **46.8** | **38.8** | **29.8** | 19.5 | 63.7 | **58.4** | **52.4** | **46.4** | 29.5 |

# Results: Long-term Adaptation (LTA) & out-of-domain warm-up

Table 2: **Adaptation at Test-time under Long-term Adaptation and with Ego4D warm-up. LTA.** We showcase the results of MiDL under the assumption that the stream of data is very long. We use unlabeled data to simulate a longer stream and report results on the validation set of each dataset. Our MiDl benefits from long-term adaptation, especially at higher missing rates (>75%). **Ego4D warm-up.** We show another use cause of MiDL, in which the assumption is having access to out-of-domain unlabeled data to adapt before deployment. The results showcase MiDL's capabilities on leveraging unlabeled-out-of-domain data to combat missing modalities.

| $1 - p_{AV}$ / Model | Epic Sounds (%) | | | | | Epic Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Baseline | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | **63.9** | 55.5 | 46.8 | 37.9 | 29.5 |
| +MiDl | 55.0 | 46.8 | 38.8 | 29.8 | 19.5 | 63.7 | **58.4** | **52.4** | 46.4 | 29.5 |
| + MiDl - LTA | 54.9 | **46.8** | **39.5** | **32.6** | **26.0** | 63.7 | **58.4** | **52.4** | **46.7** | **41.4** |
| + Ego4D Warm-up | 55.0 | 46.5 | 38.6 | 30.4 | 20.4 | 63.7 | **58.4** | **52.4** | **46.7** | 37.8 |

# Results: Long-term Adaptation (LTA) & out-of-domain warm-up

Table 2: **Adaptation at Test-time under Long-term Adaptation and with Ego4D warm-up. LTA.** We showcase the results of MiDL under the assumption that the stream of data is very long. We use unlabeled data to simulate a longer stream and report results on the validation set of each dataset. Our MiDl benefits from long-term adaptation, especially at higher missing rates (>75%). **Ego4D warm-up.** We show another use cause of MiDL, in which the assumption is having access to out-of-domain unlabeled data to adapt before deployment. The results showcase MiDL's capabilities on leveraging unlabeled-out-of-domain data to combat missing modalities.

| $1 - p_{AV}$ / Model | Epic Sounds (%) | | | | | Epic Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Baseline | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | **63.9** | 55.5 | 46.8 | 37.9 | 29.5 |
| +MiDl | 55.0 | 46.8 | 38.8 | 29.8 | 19.5 | 63.7 | **58.4** | **52.4** | 46.4 | 29.5 |
| + MiDl - LTA | 54.9 | **46.8** | **39.5** | **32.6** | **26.0** | 63.7 | **58.4** | **52.4** | **46.7** | **41.4** |
| + Ego4D Warm-up | 55.0 | 46.5 | 38.6 | 30.4 | 20.4 | 63.7 | **58.4** | **52.4** | **46.7** | 37.8 |

# Results: Long-term Adaptation (LTA) & out-of-domain warm-up

Table 2: **Adaptation at Test-time under Long-term Adaptation and with Ego4D warm-up. LTA.** We showcase the results of MiDL under the assumption that the stream of data is very long. We use unlabeled data to simulate a longer stream and report results on the validation set of each dataset. Our MiDl benefits from long-term adaptation, especially at higher missing rates (>75%). **Ego4D warm-up.** We show another use cause of MiDL, in which the assumption is having access to out-of-domain unlabeled data to adapt before deployment. The results showcase MiDL's capabilities on leveraging unlabeled-out-of-domain data to combat missing modalities.

| $1 - p_{AV}$ / Model | Epic Sounds (%) | | | | | Epic Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Baseline | **55.1** | 45.6 | 37.1 | 28.3 | 19.5 | **63.9** | 55.5 | 46.8 | 37.9 | 29.5 |
| +MiDl | 55.0 | 46.8 | 38.8 | 29.8 | 19.5 | 63.7 | **58.4** | **52.4** | 46.4 | 29.5 |
| + MiDl - LTA | 54.9 | **46.8** | **39.5** | **32.6** | **26.0** | 63.7 | **58.4** | **52.4** | **46.7** | **41.4** |
| + Ego4D Warm-up | 55.0 | 46.5 | 38.6 | 30.4 | 20.4 | 63.7 | **58.4** | **52.4** | **46.7** | 37.8 |

# Results: Other missing modalities

Table 4: **Adaptation at Test-time - Other Missing Modalities**. In this table we show the results using the complementary modality for each of the datset, *i.e.* video for Epic Sounds and Audio for Epic Kitchens. We observe that MiDL improves consistently under this setup, highlighting its robustness to different types of modalities missing at test time.

| $1 - p_{AV}$ / Model | Epic Sounds (%) | | | | | Epic Kitchens (%) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 25 | 50 | 75 | 100 | 0 | 25 | 50 | 75 | 100 |
| Unimodal | 46.5 | 46.5 | 46.5 | 46.5 | 46.5 | 63.2 | 63.2 | 63.2 | 63.2 | 63.2 |
| Baseline | **55.1** | 53.4 | 51.8 | 50.5 | 48.8 | **63.9** | 61.0 | 58.0 | 54.8 | 52.1 |
| +MiDl | 55.0 | 53.3 | 51.8 | 50.7 | 48.8 | 63.7 | 61.6 | 59.1 | 55.9 | 52.1 |
| +MiDl - LTA | 55.0 | **53.4** | **52.0** | **51.0** | **49.4** | 63.7 | **61.7** | **59.5** | **57.3** | **55.3** |

# Results: Different pre-training & architecture

Table 3: **MiDL performance with self-attention baseline.** We showcase the effectiveness of MiDL with multimodal self-attention. MiDL enhances performance across all missing rates, underscoring its robustness and adaptability to various underlying architectures.

| Model \ $1 - p_{AV}$ | Epic-Sounds (%) 0 | 25 | 50 | 75 | 100 |
|---|---|---|---|---|---|
| Self-Att. Baseline | 45.3 | 38.8 | 32.7 | 26.7 | 20.5 |
| +Shot | 45.5 | 39.0 | 32.8 | 26.8 | 20.7 |
| +Tent | 45.3 | 38.6 | 32.3 | 26.0 | 19.8 |
| +ETA | 45.3 | 38.8 | 32.7 | 26.7 | 20.5 |
| +MiDl (ours) | **45.5** | 39.5 | 33.8 | 27.5 | 20.5 |
| +MiDl - LTA (ours) | **45.5** | **39.6** | **34.5** | **29.0** | **23.2** |

Table 5: **MiDl performance with Omnivore pretraining.** MiDl is highly effective when applied to Omnivore model, demonstrating its effectiveness with a different pretraining strategy.

| Model \ $1 - p_{AV}$ | Epic-Kitchens (%) 0 | 25 | 50 | 75 | 100 |
|---|---|---|---|---|---|
| Omnivore Baseline | **65.6** | 48.1 | 47.6 | 46.0 | **44.2** |
| +MiDl (ours) | **65.6** | **57.6** | **52.4** | **47.5** | **44.2** |

# Results: Loss Components

Table 6: **Analyzing MiDl components.** We analyze the different components of MiDL. When the Mutual-Information (MI) component is missing, the model does not have any reason to adapt since the KL divergence is maximized by predicting the same as the base model. When KL is not present, the MI alone deviates from the initial results and performs poorly under higher missing rates.

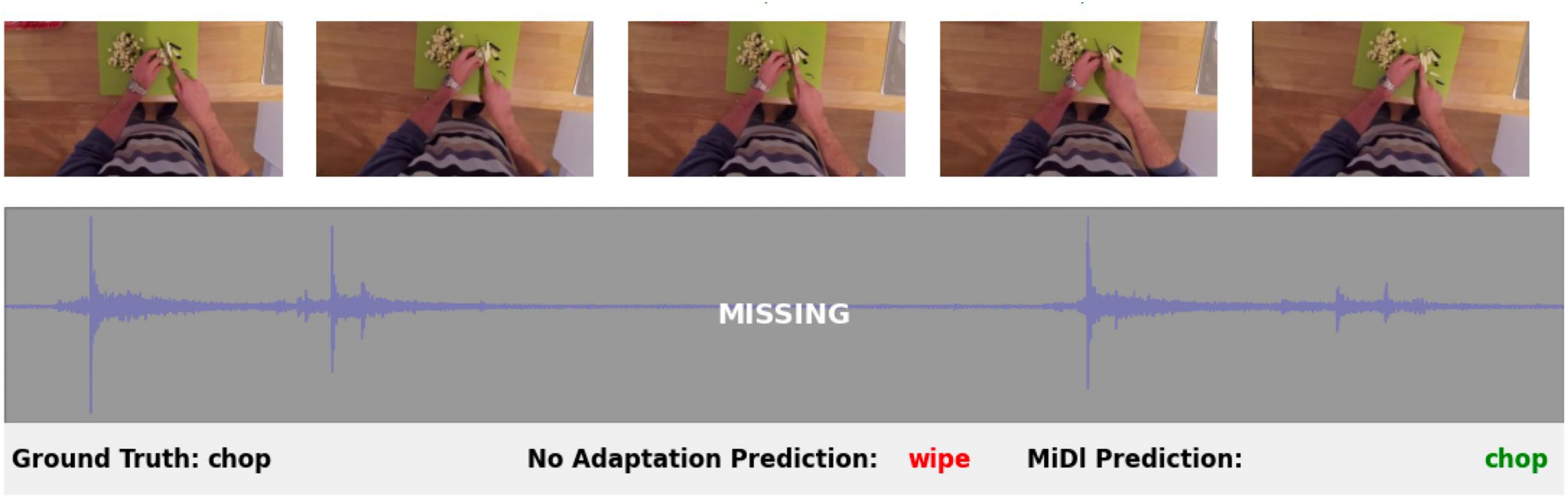
| Model \ $1 - p_{AV}$ | $\mathcal{L}_{\text{MI}}$ | $\mathcal{L}_{\text{KL}}$ | Epic Sounds (%) 0 | 25 | 50 | 75 | Epic Kitchens (%) 0 | 25 | 50 | 75 |
|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | ✗ | ✗ | 55.1 | 45.6 | 37.1 | 28.3 | **63.9** | 55.5 | 46.8 | 37.9 |
| + Dl | ✗ | ✓ | **55.2** | 45.6 | 37.1 | 28.3 | 63.9 | 55.5 | 46.8 | 37.9 |
| + Mi | ✓ | ✗ | 40.4 | 39.3 | 36.1 | 29.6 | 53.5 | 50.5 | 47.6 | 45.9 |
| +MiDl (ours) | ✓ | ✓ | 55.0 | **46.8** | **38.8** | **29.8** | 63.7 | **58.4** | **52.4** | **46.4** |

# Results: Loss Components

Table 6: **Analyzing MiDl components.** We analyze the different components of MiDL. When the Mutual-Information (MI) component is missing, the model does not have any reason to adapt since the KL divergence is maximized by predicting the same as the base model. When KL is not present, the MI alone deviates from the initial results and performs poorly under higher missing rates.

| Model \ $1 - p_{AV}$ | $\mathcal{L}_{\text{MI}}$ | $\mathcal{L}_{\text{KL}}$ | Epic Sounds (%) | | | | Epic Kitchens (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 25 | 50 | 75 | 0 | 25 | 50 | 75 |
| Baseline | ✗ | ✗ | 55.1 | 45.6 | 37.1 | 28.3 | **63.9** | 55.5 | 46.8 | 37.9 |
| + Dl | ✗ | ✓ | **55.2** | 45.6 | 37.1 | 28.3 | 63.9 | 55.5 | 46.8 | 37.9 |
| + Mi | ✓ | ✗ | 40.4 | 39.3 | 36.1 | 29.6 | 53.5 | 50.5 | 47.6 | 45.9 |
| +MiDl (ours) | ✓ | ✓ | 55.0 | **46.8** | **38.8** | **29.8** | 63.7 | **58.4** | **52.4** | **46.4** |

# Challenges of missing modalities in multimodal learning



Thank you