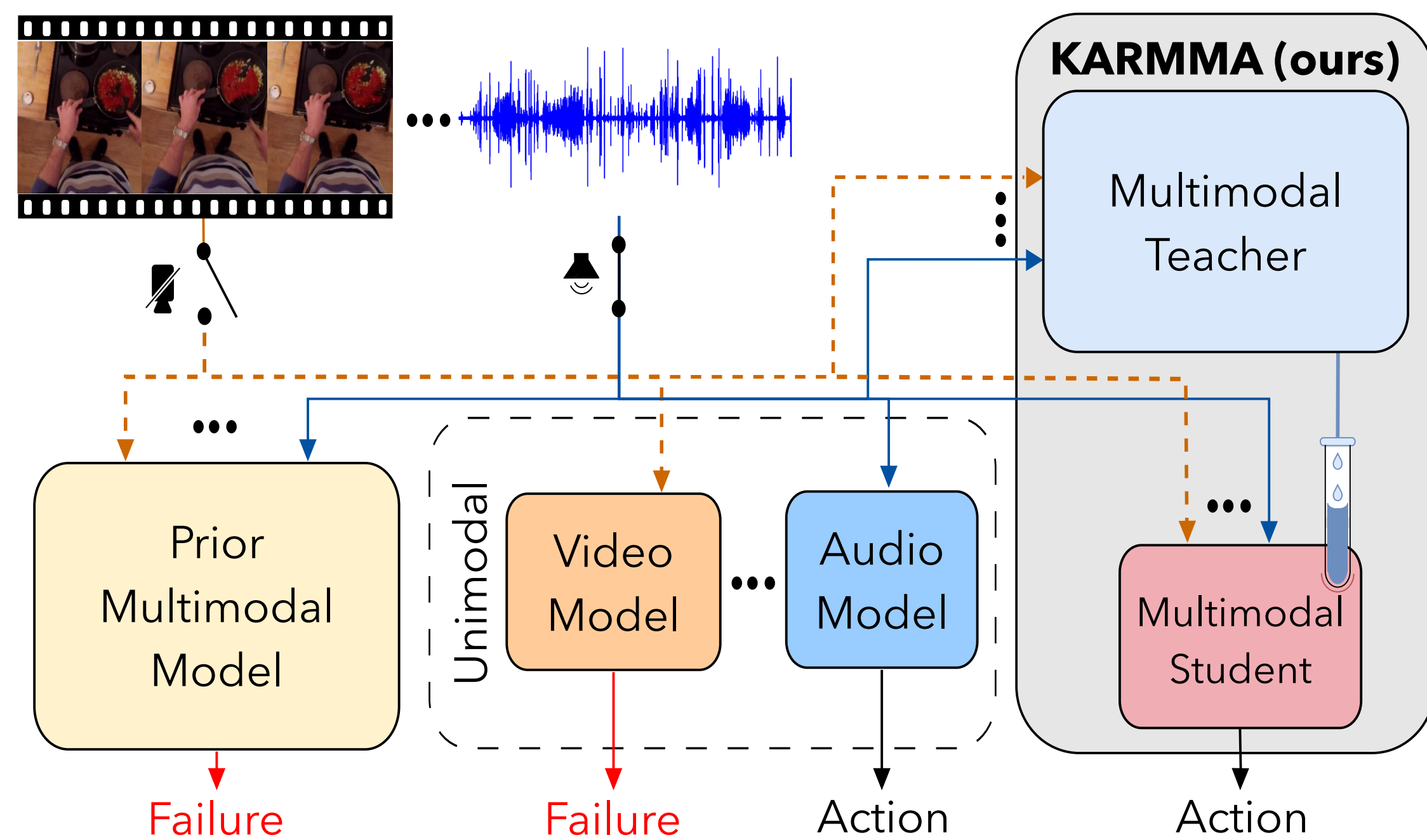


TL;DR

KARMMA is a **multimodal-to-multimodal** distillation framework for **egocentric action recognition** that **does not require modality-aligned data** and supports **any subset of modalities at inference**. It produces a **lightweight student robust to missing modalities** without retraining.

Overview

Motivation:



Goal: Train with incomplete modalities and enable efficient inference with any available subset of trained modalities.

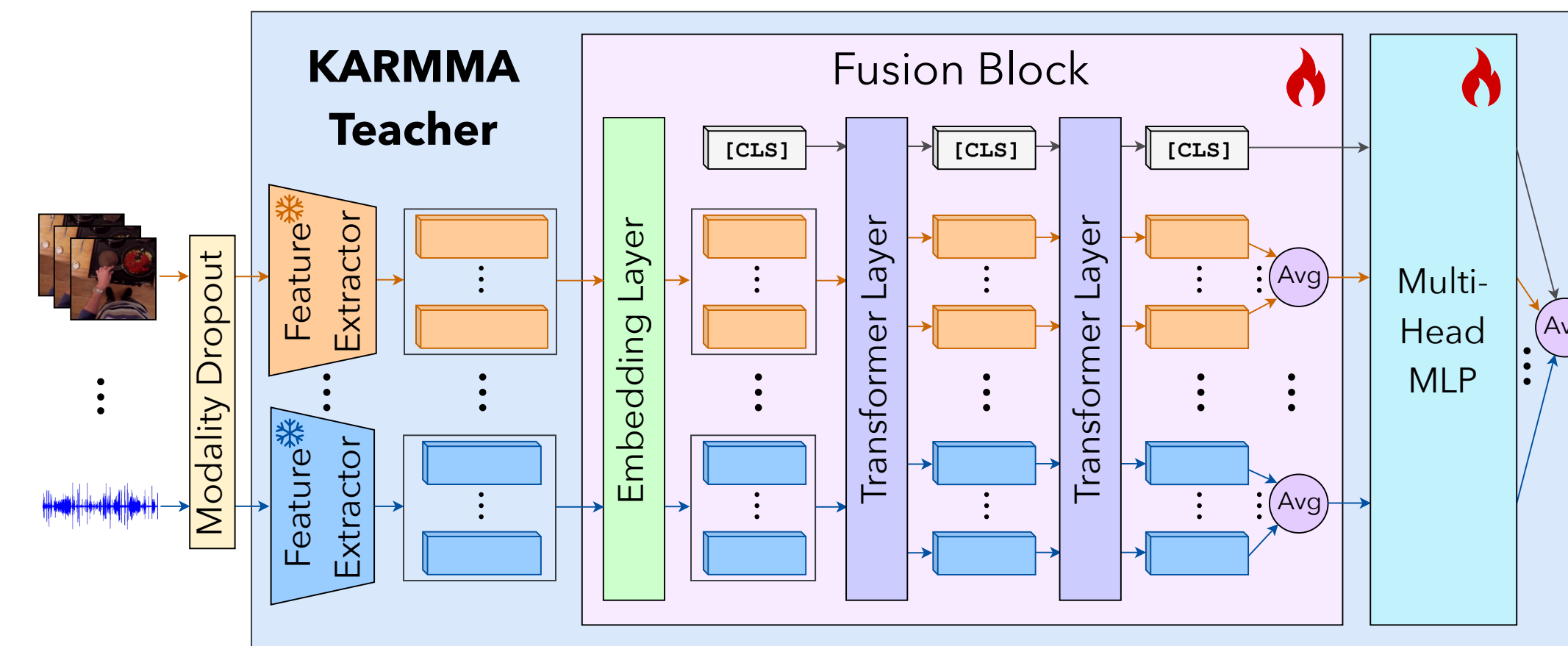
References & Acknowledgments

- [1] Dima Damen et al. Scaling egocentric vision: The EPIC-KITCHENS dataset. In ECCV, 2018.
- [2] Raghav Goyal et al. The "something something" video database for learning and evaluating visual common sense. In ICCV, 2017.

This work was funded by DFG EXC-3057/1 "Reasonable Artificial Intelligence" (Project No. 533677015), and supported by hessian.AI. SSM has been funded by DFG Grant No. 529680848. MSV, APY, JBC, and JJG further acknowledge support by MCIN/AEI/FEDER projects PID2024-158322OB-I00 and PID2021-125209OB-I00, JIUZ2024-IyA-07, DGA 2022-2026 scholarship, and SMT Erasmus+ Grant 2022-1-ES01-KA131-HED-000065592.

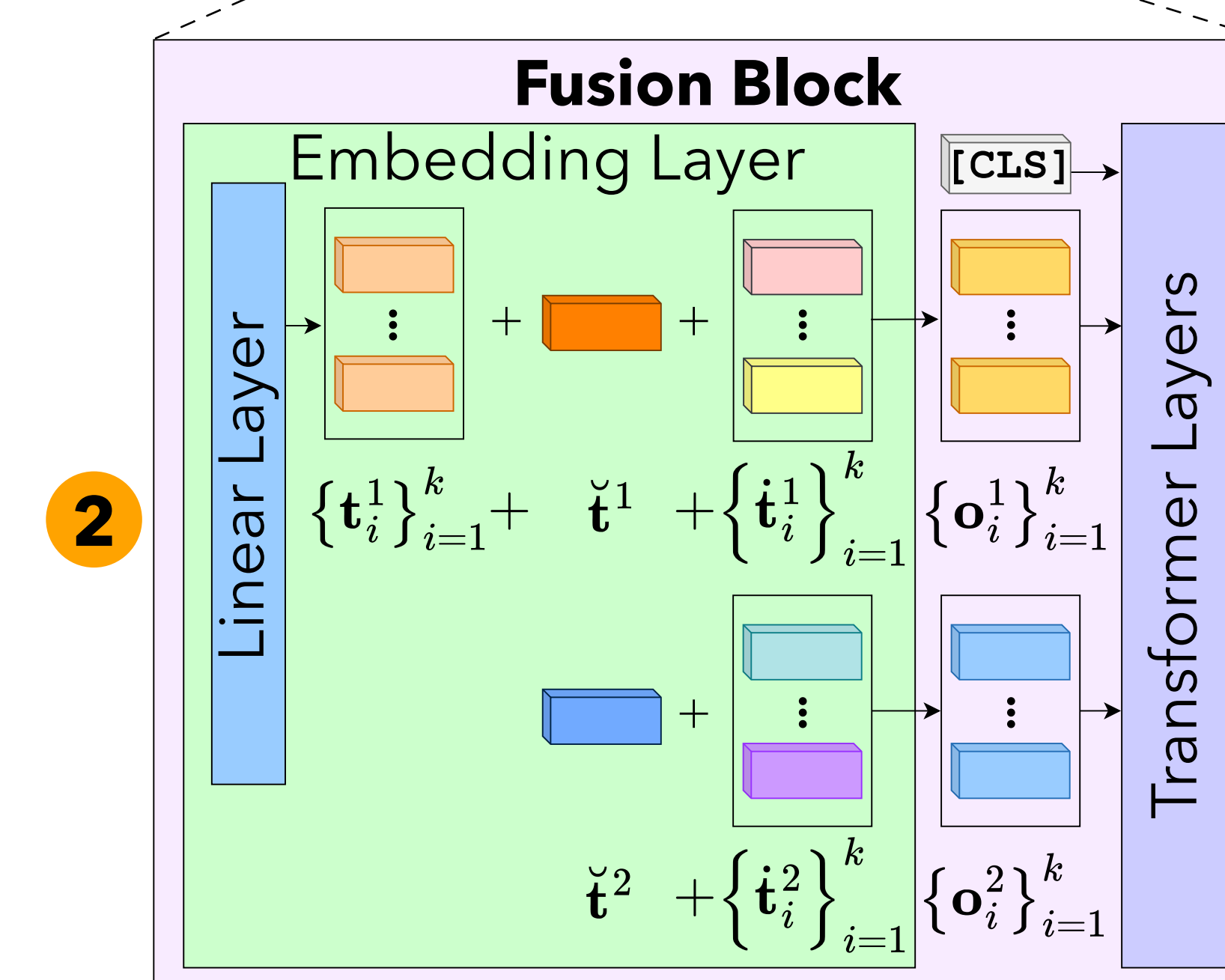
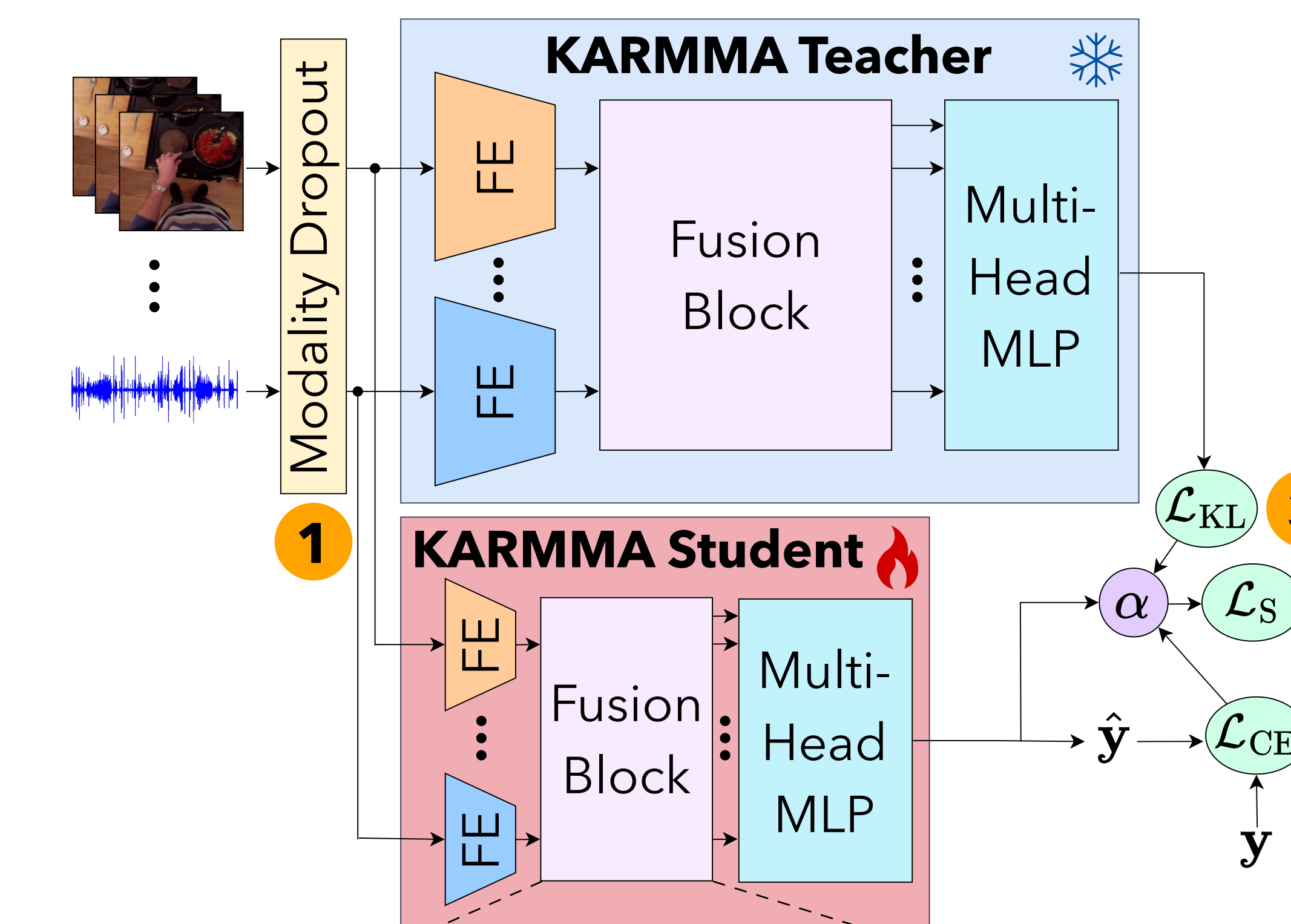
KARMMA

1 Multimodal teacher training



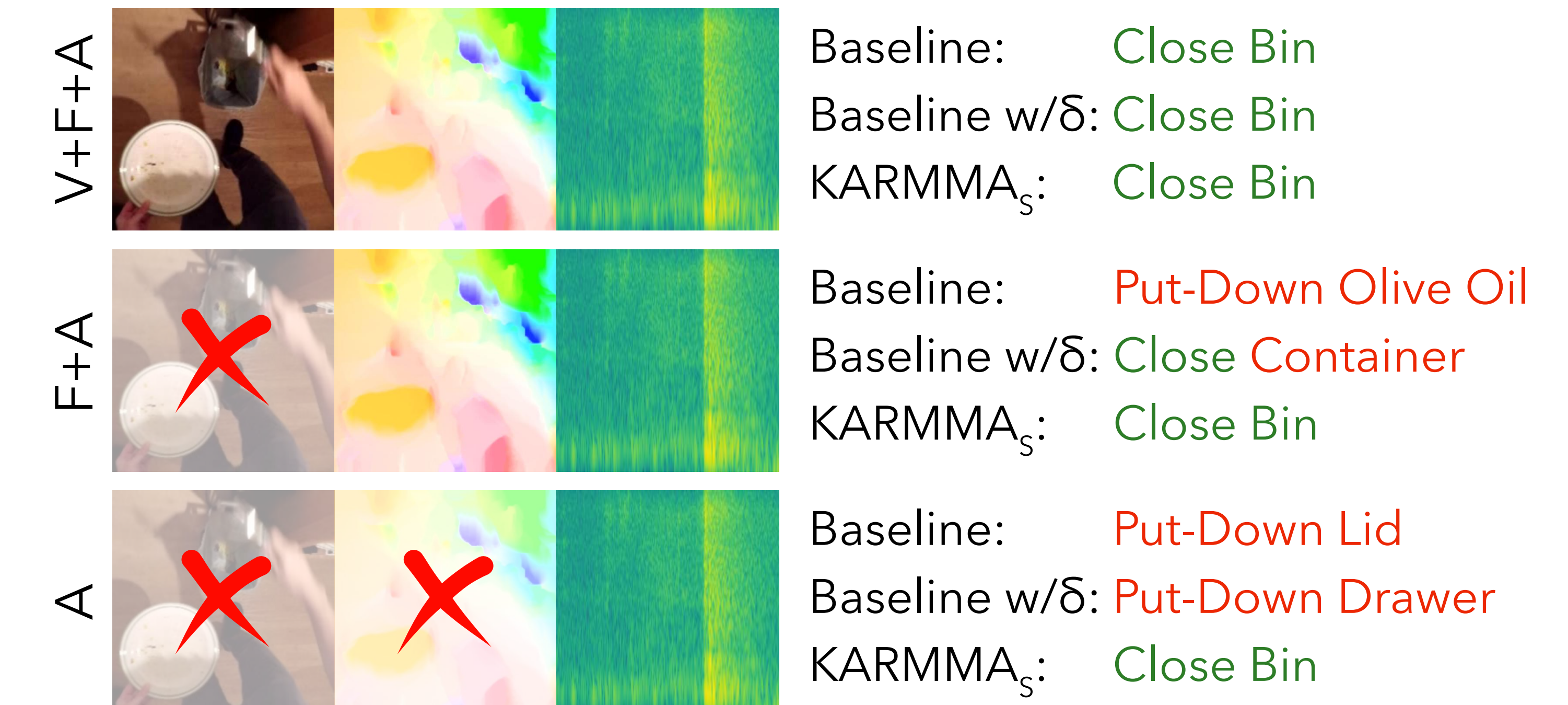
2 Multimodal student with KARMMA enhancements:

- 1 Modality dropout
- 2 Missing modality strategy
- 3 Knowledge distillation

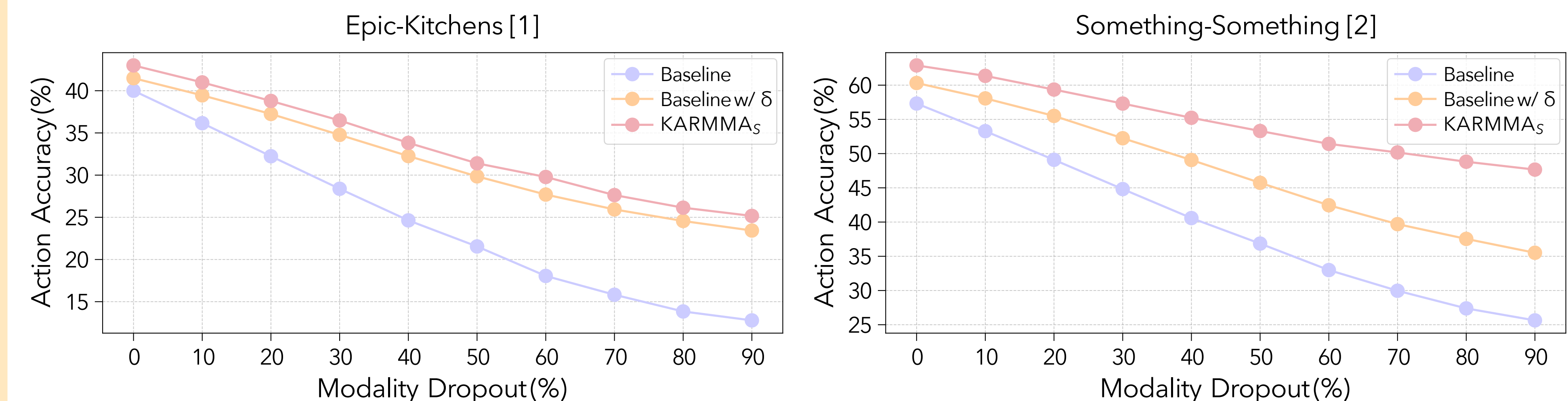


Results

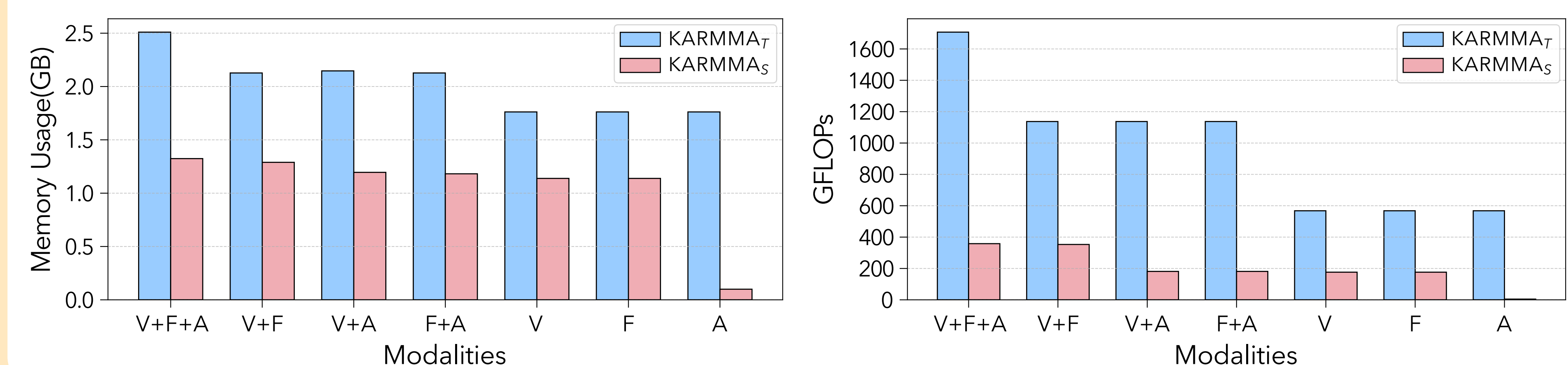
1 Qualitative results on Epic-Kitchens [1]



2 Robustness to missing modalities



3 Student resource efficiency gains over the teacher on Epic-Kitchens [1]



Takeaways

- KARMMA **does not require aligned inputs** for training
- Enables **any-to-any inference** over trained modalities
- Improves **robustness under missing-modality** conditions
- **Reduces** student **memory by ~50%** compared to the teacher

Project Page

